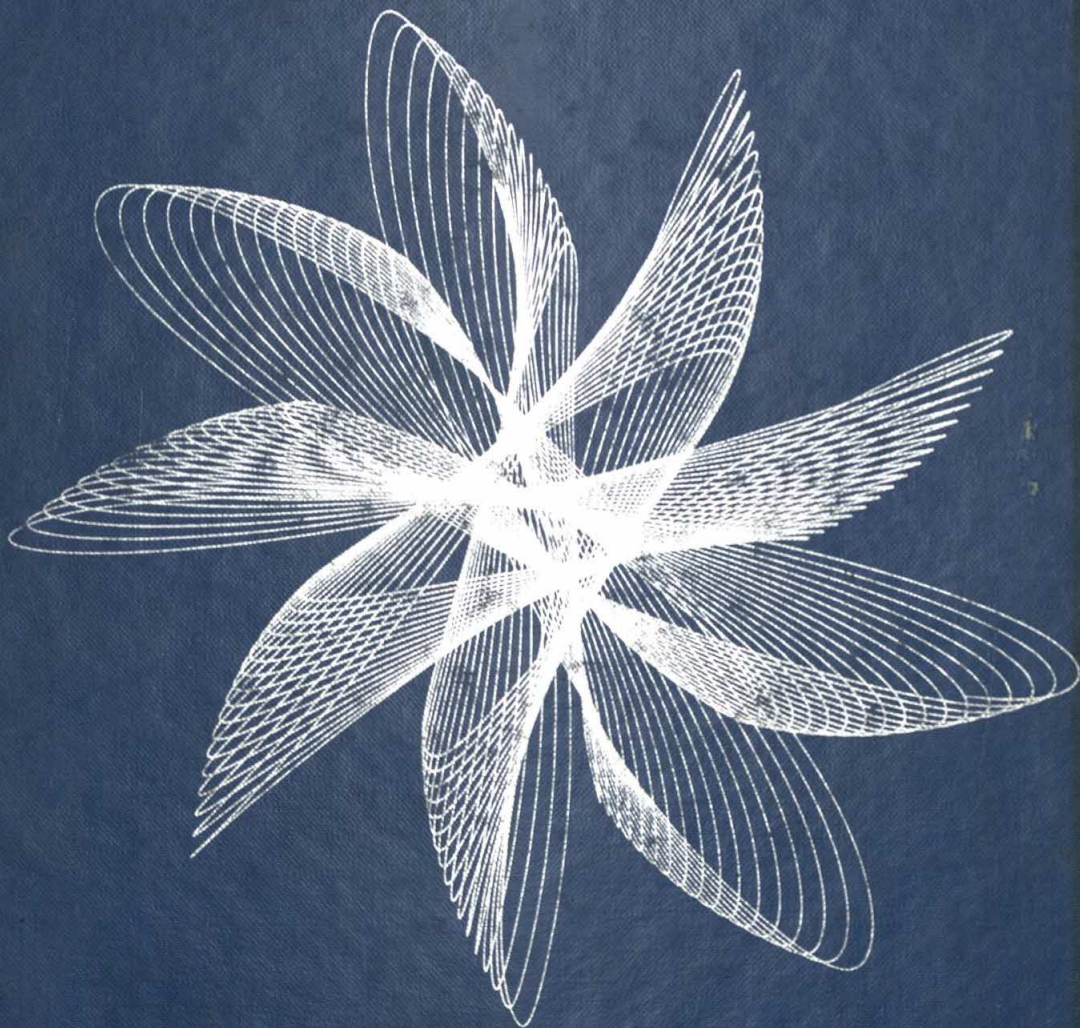# STATISTICS
A BEGINNING

KUEBLER / SMITH
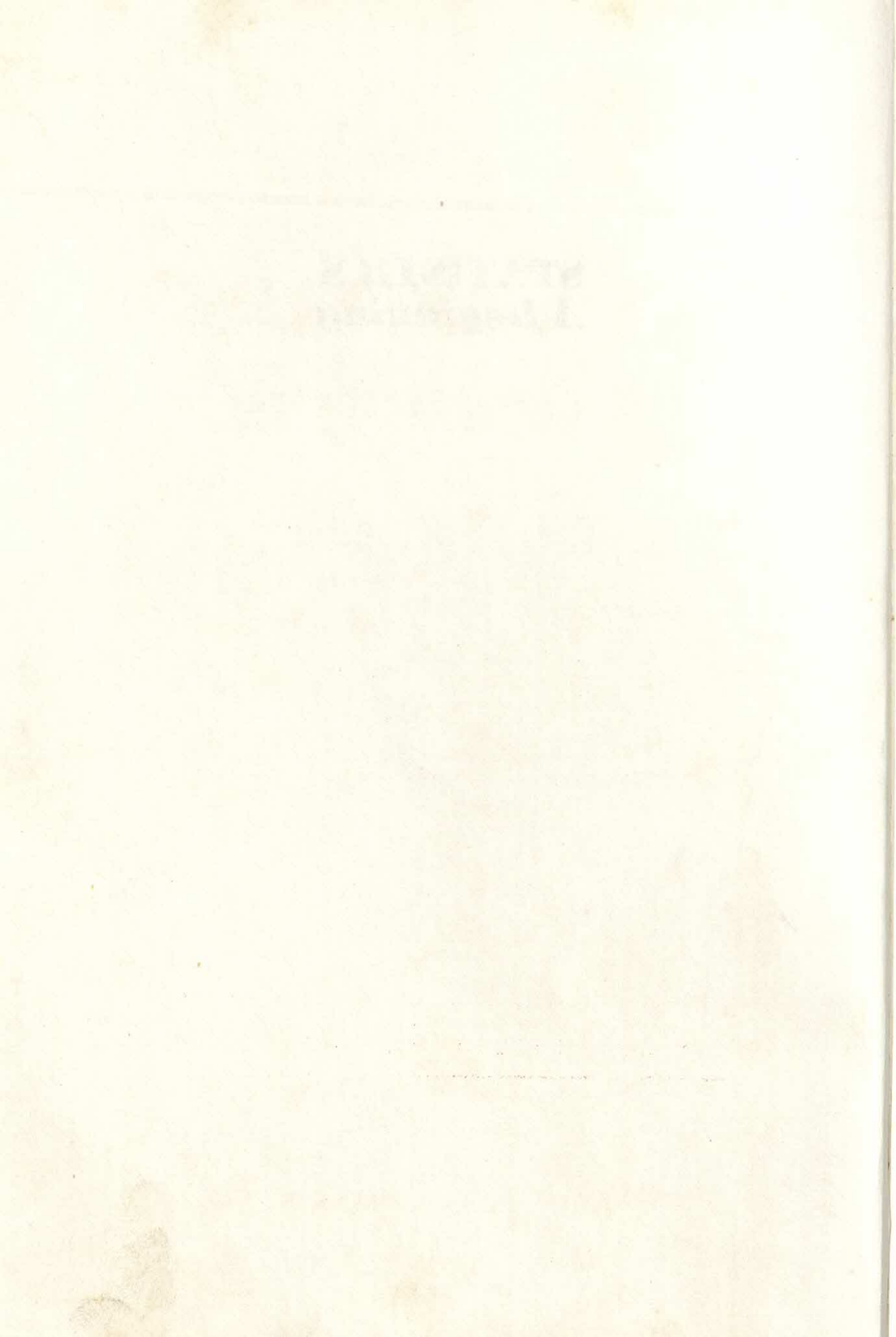
STATISTICS
A Beginning

STATISTICS
A Beginning

Roy R. Kuebler
*University of North Carolina
at Chapel Hill*

and

Harry Smith, Jr.
*Mount Sinai School of Medicine
The City University of New York*

JOHN WILEY & SONS, Inc.

Preface

This book is intended to introduce the well-tempered person to the basic notions and processes of probability and statistics as they apply to analyzing data and drawing conclusions therefrom. To be well tempered, in our view, is mostly a matter of having some modest interest in the subject, a bit of industry and patience, and a reasonable recollection of early high-school mathematics.

These days we all are besieged with numerical data and numerical arguments seeking to influence us one way or another. The cynic in us says sarcastically that "you can prove anything with statistics." The honest truth is that you can *prove* absolutely *nothing* with statistics. You *can* calculate odds, and you *can* make decisions in chancy situations with some specified control over the risk of error, and you *can* talk back to the statisticians. We think that these are worthwhile aims, and that they are attainable to a decent practical degree without elaborate training. Thus we hope for an audience of students from late high school up, technicians in and out of technical institutes, and businessmen of all varieties, indeed the general public.

The book is elementary in that it begins at the beginning of the subject and deals with just the basics. It is elementary too in mathematical level—no calculus, no set theory, and no complicated derivations. But it is a mathematics book, written by two people who are convinced that the grasp of even the most rudimentary mathematical idea requires practice of the do-it-yourself kind. So there are many examples showing mathematical details and many exercises for the reader to work on.

After much experience in studying and teaching statistics, and working as consultants with operational, management, and research workers, the authors have been led to try their hand at setting down an introduction to statistical analysis that is at once elementary and precise, intuitively motivated but also theoretically sound. It is a manual that gives a first level of statistical competence based on mathematical ideas and procedures. Though elementary and thus incomplete for advanced study, nothing in the book will have to be "unlearned" if and when the reader goes on to higher levels. Our aim may well

have exceeded our grasp, but we hope for that corrective aid that comes from irate readers who will not suffer in silence.

To all of the people with whom we have worked in statistics classes and statistics consultations we owe a debt of thanks for asking the questions and testing the answers that we have tried to organize in this book. To all of our friends who have given encouragement and advice we are indebted. High on the list must be Mrs. Doris Smith, who lived through it all and read every word, with editorial pencil in hand, to give us the critical comments of the general reader. The contents of Tables A-1, A-4, A-5 were produced by the computing group of the Department of Biostatistics, University of North Carolina at Chapel Hill. To Mrs. Karen Wendt and Mrs. Delores Gold we are most grateful for patience and skill beyond the call of duty in translating our handwriting into a typewritten manuscript ready for the printer. All of these good people deserve a share of credit for any success our efforts meet. The failures and errors that you find must of course be charged completely to the authors.

<div style="text-align: right">

Roy R. Kuebler

Harry Smith, Jr.

</div>

CHAPEL HILL, NORTH CAROLINA
NEW YORK, NEW YORK
JULY 1976

Note: A manual for the teacher and a workbook for the student are available to supplement the use of this text.

Contents

4—STATISTICS AND CHANCE

5—EDUCATED GUESSING

7—SORTING OUT THE CATEGORIES

8—PREDICTING WITH CONFIDENCE

STATISTICS
A Beginning

1

Data For Statistics

1.1 INTRODUCTION

All of us are constantly annoyed by the confusion and misunderstandings that occur in verbal and written communication. It is then reasonable to ask, "how can things be more clearly stated so that such conflicts can be avoided?" A scientist such as a chemist, mathematician, or engineer attempts to resolve these problems by becoming very quantitative. For instance, a chemist will measure the amount of liquid in a vessel to the nearest tenth of a cubic centimeter and state, "there are 70.6 cubic centimeters of liquid in this vessel." Such a statement is very clear and very seldom leads to any misunderstanding.

However, not every quantitative statement is this clear. Consider the following statements:

1. John Jones's salary is one half of Joe Emery's.
2. Mary is twice as smart as Sarah.
3. This is the fifth largest snowfall in Schodack's history.
4. In one study, Crest toothpaste reduced cavities 42 percent when compared to a control.

Each statement has some quantitative aspect that, by itself, is understandable; that is, we know the meaning of "one half," "twice," "fifth," "42 percent," and so on, but there is still much opportunity for confusion and misinterpretation.

For example, what is really

meant by statement number 2? Is there any valid way that such a judgment could be made? Are we talking about grades in school, results on achievement tests, or just some overall subjective judgment? In any case the word "twice" conveys something definite to the listener. He knows that $2 \times 4 = 8$, $2 \times 10 = 20$, and so on. However, the use of the very specific quantitative word "twice" can still be subject to a great deal of controversy.

Statement number 3 might convey something important to a weatherman in Schodack, but it doesn't convey much to anyone else. Everyone knows that there have been four snowfalls larger than this one, but there is no way for us to understand anything more than just that. Fifth relative to what? How large was the snowfall?

To make statements less quantitative doesn't help much either. For example, statements like "this is the worst smog this year" or "the policeman gave Eunice a ticket for speeding" conjure up various pictures in people's minds. Different people interpret things differently, and the more subjective or personal interpretation possible, the more confusing things become.

The subject "statistics" is an attempt to bring some understanding to the use of quantitative measures in any kind of statement one might like to make. We will find that the use of statistical quantities will be *severely* restricted by the kind of data one has, by the way in which we will draw conclusions and make predictions, and finally, by the people to whom statistical reports are submitted. Being thus restricted, and carefully defined, statistical quantities have a chance of keeping quantitative statements clear of confusion and misinterpretation.

In this book we hope to give you the beginning ideas of what *statistics* is all about. And the beginning of that beginning can well be your consideration of a number of illustrative examples of quantitative statements that leave something to be desired. For that purpose the following exercises have been compiled.

EXERCISES

1.1.1 For each of the statements below, do the following: (a) underline the numerical quantities, (b) determine what additional information is required to enable you to discuss the statement, (c) rank the statements A–F in order of clarity, 1 being the most clear and 6 being the most unclear, and (d) discuss the usefulness of each statement.

A. "Love-ins that extend for more than one night become a deadly bore, even at age 17. I think 28 minutes is roughly all the time you need for a love-in." (Peter Drucker, Our Top-Heavy Corporations, *Dun's*, April 1971.)

B. "Analyses show that 18% of staff employees are minorities and 63% women." [Powers, Mary F., End discrimination to hold onto Federal funding, *Coll. Manage.*, May 1971 (article on Univ. Pittsburgh).]

C. "In transportation, while U.S. road vehicles continue to increase in numbers twice as fast as the human population, the creation of new mass transportation systems to relieve our choking roads lags far behind—though not for lack of abundant technology." (Lessing, Lawrence, The senseless war on science, *Fortune*, March 1971, p. 154.)

D. "Surveys indicate that as many as 40% of engineers would choose a different profession if they had a chance to start again." (Gooding, Judson, The engineers are redesigning their own profession, *Fortune*, June 1971, p. 72.)

E. "Approximately 43% of the women and 30% of the men failed to recognize feelings of love until after 20 or more dates . . . ." [Brothers, J., Ask Dr. Joyce Brothers, *Durham Sun*, Friday, July 3, 1970, Durham N.C. (quotation from a Midwestern Univ. Survey).]

F. "A study of 10,000 workers in the Chicago area has found that 3000 of them can be considered high risk candidates for heart attacks . . . ." (*N.Y. Times*, March 22, 1970.)

1.1.2 Collect three examples of interesting but confusing statements from any of the media such as newspapers, magazines, or television. Explain why the statements are confusing or misleading.

1.2 DATA ON SOME MECCA COMMUNITY COLLEGE STUDENTS

Discussion of the difficulties associated with the use of quantities, numbers, percentages, and so on in Section 1.1 leads us naturally to consider the basic characteristics of data. Is there a structure to data, and, if so, will this structure help us in making more valid use of numerical or other quantifiable information?

In order to help us understand the basic structure of data, we shall use the results of surveying 180 upper-class students in a large community-college system to indicate the kinds of data we shall be considering in this text. The data obtained from each student were coded and recorded below.

TABLE 1.2.1 Survey Results on 180 Nonfreshman Community College Students (Mecca Community College)

Student Number	Sex	Commuting Distance from Home to College (Miles to Nearest One Half Mile)	Political Party Preference[a]	Marijuana Question[b]	Freshman G.P.A.[c]
1	M	24	R	3	2.42
2	M	15	D	1	1.75
3	F	9	N	4	2.06
4	M	15	R	1	3.00
5	M	25.5	O	3	2.22
6	M	9.5	R	4	2.71
7	M	11.5	N	5	3.26
8	M	5.5	R	1	2.90
9	F	5	R	1	2.00
10	M	32	D	5	3.59
11	F	24	N	5	2.46
12	M	26.5	D	4	2.33
13	M	5.5	D	5	3.15
14	M	35	R	1	2.47
15	M	5	D	1	3.11
16	M	6.5	N	4	0.81
17	M	7	D	1	1.13
18	M	22.5	D	4	2.60
19	M	32	D	4	2.84
20	F	6	R	4	2.70
21	M	10	N	4	2.26
22	M	3.5	R	3	1.35
23	F	7	N	5	2.12
24	M	4	D	4	1.54
25	M	3.5	R	2	1.61
26	M	28.5	R	1	2.15
27	M	10	D	1	3.18
28	M	32	N	1	2.66

Student Number	Sex	Commuting Distance from Home to College (Miles to Nearest One Half Mile)	Political Party Preference[a]	Marijuana Question[b]	Freshman G.P.A.[c]
29	M	7.5	D	4	1.34
30	M	29.5	N	4	2.57
31	M	4	R	4	1.71
32	F	15	D	4	2.41
33	F	3.5	D	1	3.03
34	M	4	N	4	2.50
35	F	27.5	D	5	2.91
36	M	2	N	1	1.20
37	F	4.5	N	1	2.23
38	M	36.5	D	4	3.09
39	M	4	O	2	1.66
40	F	3	R	2	2.67
41	F	16.5	R	4	2.48
42	F	15	D	4	3.06
43	M	56.5	D	2	3.31
44	F	27.5	N	5	2.39
45	M	25	R	2	2.71
46	F	13.5	R	1	2.74
47	M	13.5	R	2	2.85
48	F	20	N	4	2.69
49	M	56.5	D	4	3.37
50	M	2	N	1	2.63
51	M	12.5	R	2	2.09
52	F	0.5	N	5	2.71
53	F	3.5	N	1	2.68
54	M	3.5	N	4	2.68
55	F	7	D	1	2.00
56	M	11	N	4	2.89
57	M	0.5	N	4	1.94
58	F	5.5	D	3	2.59
59	F	12.5	R	4	2.92
60	F	26.5	D	4	2.08
61	F	14.5	R	1	2.62
62	F	0.5	R	4	2.18
63	F	28.5	D	4	2.35
64	M	7.5	N	1	1.06
65	M	11.5	N	5	2.57
66	F	20.5	D	1	2.23
67	M	1	N	4	1.91
68	F	25.5	D	4	2.98
69	F	13	R	2	2.41
70	F	17.5	D	5	2.03
71	F	17	N	4	2.47
72	M	11.5	N	4	2.12
73	M	3.5	R	1	2.38
74	M	7.5	D	5	2.40
75	F	2	N	5	2.95
76	M	1	N	4	2.81

Student Number	Sex	Commuting Distance from Home to College (Miles to Nearest One Half Mile)	Political Party Preference[a]	Marijuana Question[b]	Freshman G.P.A.[c]
77	M	3	N	4	2.78
78	F	8.5	D	4	2.44
79	F	19.5	D	4	2.53
80	M	6.5	D	4	2.50
81	F	2.5	R	1	1.64
82	M	3	N	5	2.25
83	M	1.5	N	4	2.28
84	M	29.5	R	3	2.14
85	M	15.5	N	5	3.10
86	F	13.5	R	1	3.36
87	M	12	D	1	2.90
88	M	3	N	4	1.46
89	M	4	D	5	3.36
90	F	23.5	N	1	2.34
91	M	28.5	N	4	2.12
92	M	3	O	1	3.05
93	M	20	R	4	2.19
94	M	17	N	4	3.00
95	M	0.5	R	4	2.46
96	M	3.5	N	3	3.22
97	F	6	D	4	3.16
98	M	19.5	O	4	2.04
99	M	1	R	5	2.21
100	F	1	R	1	3.21
101	M	35	D	4	1.83
102	M	1	D	5	4.00
103	M	13	R	4	1.56
104	F	13	N	4	2.39
105	F	10.5	R	4	2.40
106	F	24	D	1	2.61
107	M	22.5	D	4	2.28
108	M	24	D	5	3.17
109	M	15.5	D	2	2.73
110	M	6	N	4	3.09
111	M	19.5	D	4	2.78
112	M	3	D	4	1.67
113	F	1	D	1	2.35
114	M	1	N	4	1.98
115	F	8	D	1	1.82
116	M	7	D	4	1.72
117	M	21.5	N	5	2.85
118	M	8	N	4	2.12
119	M	3	N	4	1.71
120	F	2.5	N	1	3.19
121	F	6.5	N	1	3.86
122	M	5	D	1	2.51
123	M	46.5	N	5	2.08
124	M	2	R	4	1.88

Student Number	Sex	Commuting Distance from Home to College (Miles to Nearest One Half Mile)	Political Party Preference[a]	Marijuana Question[b]	Freshman G.P.A.[c]
125	F	22.5	D	1	2.13
126	F	12.5	N	5	1.94
127	F	17	D	1	1.75
128	F	6	R	2	1.53
129	M	25.5	N	5	1.67
130	M	5	R	4	3.55
131	F	19	R	4	2.16
132	M	6	R	4	2.37
133	F	5	N	4	2.42
134	F	36.5	N	5	2.44
135	F	9.5	R	4	2.02
136	M	10.5	R	4	2.92
137	M	36.5	D	5	2.37
138	M	15	R	1	2.72
139	F	3.5	R	1	3.09
140	F	9	R	1	3.09
141	M	3.5	O	4	2.61
142	F	12	N	5	2.26
143	F	8	R	1	2.15
144	F	21.5	O	4	2.08
145	M	7.5	N	3	2.59
146	M	4.5	R	4	2.40
147	F	23.5	D	4	1.80
148	M	12	N	5	2.25
149	M	7.5	N	1	1.93
150	M	7	R	1	1.27
151	F	1.5	D	4	3.40
152	M	25	R	4	2.70
153	M	4	D	2	1.91
154	F	6	D	3	2.13
155	F	5	N	1	3.29
156	F	7	N	4	2.91
157	M	10	N	2	2.15
158	F	4	D	4	2.17
159	M	29.5	R	4	2.79
160	F	28.5	D	1	2.36
161	F	27.5	N	4	2.60
162	F	4.5	N	4	2.00
163	F	12	D	1	2.50
164	F	0.5	R	4	2.10
165	F	12.5	N	5	2.66
166	F	4	D	1	1.79
167	F	4	D	4	1.86
168	M	6	D	1	2.18
169	M	8	R	4	2.63
170	F	7	N	5	1.51
171	F	9.5	R	4	1.85
172	M	20.5	D	5	1.00

Student Number	Sex	Commuting Distance from Home to College (Miles to Nearest One Half Mile)	Political Party Preference[a]	Marijuana Question[b]	Freshman G.P.A.[c]
173	M	20.5	N	5	1.97
174	F	3	N	1	2.60
175	M	2.5	N	5	2.66
176	F	10.5	N	2	2.71
177	F	0.5	N	2	2.05
178	M	1	D	4	1.61
179	M	12	D	1	2.28
180	M	4	R	1	2.07

[a] Political party preference:
  R = Republican
  D = Democrat
  O = Other party
  N = No party preference.
[b] Marijuana Question: Statement: "Marijuana should be legalized"

Opinion Scale

1 Strongly Disagree	2 Mildly Disagree	3 No Opinion	4 Mildly Agree	5 Strongly Agree

[c] G.P.A. = Gradepoint average at end of freshman year (range 0–4).

We shall refer to these *data* as *observations* taken on each of five survey questions from 180 community college students. Since the total number of upper-class students (in the system) was around 10,000, these 180 students represent a small subset of the total. We shall call this subset a *sample* of the upper-class students and refer to the number of students in the subset as the *sample size*. We denote the sample size with the letter n. Thus we say that the sample size here is $n = 180$.

In case you are thinking about visiting Mecca Community College, we have to confess to you that it exists only within the pages of this book. It is a composite of facts and features of many institutions, real and imagined, put together to represent lots of places without running any danger of favoring or offending anyone.

The data in the table are of the various kinds that we meet most often in quantitative investigations. There are purely *categorical* data: male, female; Democrat, Republican, other party, or no party preference. There are data that are categorical but subject to *ranking*: strongly disagree, mildly disagree, no opinion, mildly agree, or strongly agree. And there are the numerical data with which we are most familiar in *measurement* terms: the distance between home and college, the student's gradepoint average (G.P.A.).

It would be useful for the reader to think a bit about what might come out of this set of data, sensible or otherwise. The following exercises suggest some lines of thought.

EXERCISES

1.2.1 Using the data on the 180 upper-class students at Mecca College, write five statements that you believe to be quantitatively clear.

1.2.2 Write five statements using numbers, percentages, and so on from these data that you would consider controversial.

1.2.3 What aspects of the data need clarification?

1.2.4 If you were going to gather data (similar to these) at your school, what changes in topics, categories, or procedures would you make?

1.3 DATA CLASSIFICATION

Having discussed the confusion and misunderstanding that occur when one interprets numbers and having recognized the difficulties that arise even when data are clearly structured as in the Mecca Community College data, it is nice to know that there really is a structure to data. This structure will be very useful in helping us to understand how data can be used effectively.

For the purposes of this text, we shall identify observations according to the following classification scheme: (a) discrete (nominal; ordinal) and (b) continuous (interval; ratio).

A. Discrete Data

When data are classified into categories such that any observation can fall into one and only one category, the data are said to be *discrete*. For example, in Table 1.2.1, sex, party preference, and the opinion on marijuana are examples of "discrete" data. However, discrete data can again be separated into two distinct groups: one group which is strictly qualitative in character and is called *nominal*, and another group, still qualitative, but which also has an inherent ordering characteristic and is called *ordinal*.

Nominal Observations. The first and simplest form of measurement one can use on experimental results is to classify them by some nominal attribute, for example, "this is a male," or "this student is a Democrat." The chief defining characteristics in nominal observations are their qualitative nature and the equality of status within each nominal classification. Thus decisions on the political preference of the 180 students in the sample make no allowance for different strengths of preference among those stating a preference for Democrats. Likewise if one categorized color TV sets into two mutually exclusive categories, acceptable or nonacceptable, all TV sets classified as acceptable would have equal status within the classification even though some may be merely satisfactory and others excellent. Observations of this type are classified

in a very unsophisticated manner, and treatment of the data thus obtained is limited to using the total numbers of observations in the various categories. However, obtaining this kind of data requires very little effort and is usually inexpensive and quick. Also there are times when the counts in various categories are exactly what we are looking for.

Ordinal Observations. The next step in classifying observations in a more quantitative way is to use an ordinal classification scheme. For example, bar soaps might be graded excellent, good, satisfactory, poor, and bad. While these are nominal-type classifications, they have an inherent ordering among them; that is, an *excellent* bar is in some qualitative way a better bar than one that falls in the *good* category. Likewise the opinion scale for considering the legalization of marijuana is inherently ordered. It is easy to see that this type of data can be of more use to an experimenter than can nominal observations.

Another example of this kind of data would be the following data on the classification of coal miners according to the seriousness of a lung disease called *pneumoconiosis.*

Degree of Pneumoconiosis Among 100 Coal Miners			
No Evidence	Minor Case	Major Case	Serious Case
80	12	6	2

While these data reflect an ordering of severity, no decision can be made about the magnitude of the difference between a minor case and a major case, or between a major case and a serious case. All we know is that a serious case is somehow worse than a major case and that a major case is qualitatively worse than a minor case. While such inherent ordering assists in making decisions or judgments somewhat more precise, people are just not content to make identifications, comparisons, and judgments using simple qualitative data. They want and continually strive to be more specific or precise in every judgment. The next step is to consider *distance* between categories as being more informative than what is shown with just category identification. Thus distance between categories should have a numerically meaningful value.

Only one kind of ordinal data has such distance between categories naturally. That kind is *counting* data. Toss six coins in the air, let them fall to rest, and count the number of heads. That number has to be either 0, 1, 2, 3, 4, 5, or 6. Our observations in repetitions of such tosses will thus fall into the categories 0, 1, 2, 3, 4, 5, and 6, and those categories have all the numerical rights of the numbers involved: 5 is 2 more than 3, 4 is twice as large as 2, and so on.

Such counting categories can go on and on. Set up an experiment in which you toss one coin repeatedly until a head appears. Record the number of tosses required. Well, that number can be either 1 (head on the first toss *can* occur), 2, 3, 4, or on and on. In theory, you may keep getting a tail virtually forever. Thus in observations on such an experiment, the categories of those observations are 1, 2, 3, 4, 5, and so on. Mathematicians write 1, 2, 3, 4, 5, . . . and mean by those three dots "on and on in this manner without end." We say the number of categories is *countably infinite*.

Categories of count do not have to march 0, 1, 2, 3, 4, and so on. *Pairs* of socks get counted 1, 2, 3, and onward but the corresponding *socks* get counted 2, 4, 6, and onward. Integers are 1, 2, 3, 4, . . . but *squares* of integers are 1, 4, 9, 16, ... The distinctive feature of count data, as with all ordinal data, is the discreteness of the various possible categories.

Ordinal data that are *not* counting data are often coded numerically for convenience to computers. We can code *male* as 1, *female* as 2 (or vice versa!). We can designate 1 as Republican, 2 as Democrat, 3 as Other Party, and 4 as no party preference. We *have* coded opinion on legalizing marijuana, taking 1, 2, 3, 4, and 5 in order along an opinion scale ranging between *strongly disagree* and *strongly agree*.

A hazard in using such coded identification of categories is that the codes will be interpreted as *count* data, with *distance* mistakenly assigned to the spacings between categories. For example, in the scale (or coding) used for the opinion on legalizing marijuana, the following implication might be made: With a score of 1 for strongly disagree and a score of 5 for strongly agree, then strongly agree is five times the score of strongly disagree. Is that what the coder or experimenter wanted to imply? It also implies that you can plot the data in a scale having equal spacings between categories; that is,

```
  _____
  1       2       3       4       5
```
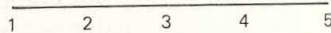
Perhaps this is not meant either. We need to be carefully on guard as to how different interpretations can be made using the same data when arbitrary scaling or coding is used. All of this leads us to search enthusiastically for data that *do* allow us to measure distances and make neat numerical comparisons. And so we go to our next classification.

B. Continuous Data

When the data are obtained using a continuous or noninterrupted scale of measurement such that numerically equal differences stand for empirically equal differences, we say that the data are continuous. There are two different types of scale for such data.

Interval-scale Data. If the scale has an arbitrary zero point, then the data are obtained using an *interval* scale. Typical examples of interval scale data are those obtained using the centigrade or Fahrenheit temperature scale, using the year designations such as 1975 and 1976, or using indices in which the zero point is determined by an arbitrary definition, as for I.Q., health status, and the like.

The following table indicates a typical ordered qualitative classification of data and also the use of an interval scale. [Pratt, Lois, The relationship of socioeconomic status to health, *Am. J. Public Health* 61 (2), 281–291 (1971), Table 1]:

TABLE 1.3.1 Quality of Health Maintenance Practices in Relation to Level of Health and Extent of Health Problems*

Quality of Health-Maintenance Practices	Average Scores on Indexes of	
	Level of Health	Extent of Health Problems
Poor	2.3	33.7
Medium	2.8	27.9
Good	3.1	24.2
Total Group	2.8	27.8

Consider the index on the status of health. Does a zero on this scale mean "no" health at all? Certainly not. The zero is arbitrarily defined. It has no real health meaning. Does the index score 3 mean "twice as healthy" as index score 1.5? Surely not; we cannot make health sense of such ratios. While such restrictions place very little constraint on the numerical analysis of these data, there are some difficulties in interpretation associated with interval-scale data.

In the data on the $n = 180$ upper-class students, the observations on G.P.A. are measured on an interval scale. We are all aware that a 2.00 or C in one class is not the same as a C in another class. Further, no one would agree to statements like, "John's G.P.A. is 3.00 and Jim's is 2.00; therefore, John is three halves or one and one half times as 'smart' as Jim." Thus any ratio calculated using an interval scale is suspect.

Another example is the "rub-for-suds" test for soap bars. In this test, the bar soap is handled in such a way that the number of standard rubs of a towel against the soap while wet is the inverse measure of the sudsing quality of the soap. The towel is rubbed against the bar in such a way that the towel is carried down to water in a pan beneath the bar at the end of every rub, and the soap is rinsed out before the next rub. The test is mechanized, and the rubbing of the soap bar is continued until enough soap is in solution to form true suds which will persist for 15 seconds after agitation. It is not necessary that the suds reach

a particular height or that they cover the surface of the water, but even so the number of rubs-for-suds is an inverse measure of the sudsing quality of the soap. The results are reported as the number of rubs required for a given amount of suds; normally this turns out to be about 100 rubs, and it is said that a bar requiring 200 rubs is only half as good as the first, standard bar. Conversely a bar requiring only 50 rubs to give a persistent suds is said to be twice as good as the standard bar, but is it? There is no absolute nature to the observations, since they are referred to an artificial zero standard. The zero thus assigned is not the lower limit at which the property of sudsing vanishes, and consequently the results are truly definable as lying on an interval scale.

Ratio-scale Data. Ratio scales are interval scales with the distinction that they have an absolute zero. Thus measurements like height, weight, amount of income, and level of iron in the blood stream are examples of ratio-scale data.

In one article is given the distribution by 1967 per capita personal income of 134 counties in Texas with a commodity distribution or food-stamp program in September 1968. ([Lukaczar, Moses, Lessons for the Federal effort against hunger and malnutrition from a case study, *Am. J. Public Health,* 61 (2), 259–276 (1971)]; data modified for purposes of this illustration.) Income values like 1999 stand for 1999.999 . . . , closing up the continuous scale needed to accommodate every possible per capita income value.

Table 1.3.2 is this example of a set of continuous data on a ratio scale. Another example is the data on the commuting distance between college and home in Table 1.2.1, where "zero" distance is a *real* zero value.

TABLE 1.3.2 *

Per Capita Income	Midpoint	Number of Counties
Less than $1500	750	20
$1500–1999	1750	50
2000–2499	2250	25
2500–2999	2750	23
3000–3499	3250	9
3500–3999	3750	6
4000–4499	4250	0
4500–4999	4750	0
5000–5999	5500	1
Total		134

TABLE 1.3.3   Summary: data classification

Type of Observation	Distinguishing Characteristic	Common Examples
I. Discrete	Observations are grouped into distinct classes	
A. Nominal	Distinct classes have no predetermined rank or order	Patient either has or does not have disease A
B. Ordinal	Distinct classes have predetermined rank	Patients are classified qualitatively by severity of disease; for example, degree of pneumoconiosis
II. Continuous	Observation may assume any value on a continuous scale	
A. Interval	Scale defined in terms of differences between observations. Zero point is arbitrary	Patients' temperature was recorded as 98° Fahrenheit. I.Q. measurements
B. Ratio	Scale differences represent real relationships in the items measured. Zero point represents total absence of attribute being measured	The percent of population living on farms, 1970 census. Median family income

1.4   EXERCISES

1.4.1   Classify the following observations as to: (a) continuous or discrete, and (b) nominal, ordinal, interval scale, or ratio-scale data.

A. A ball bearing has diameter 3.25 millimeters.
B. Twenty-three students attended history class.
C. A toss of 10 coins resulted in six heads and four tails.
D. Jerry was 13th in the math test.
E. The Smith family was classified as a middle-income group.
F. A man swam 50 yards in 62 seconds.
G. Relative humidity reading of 55 percent.
H. The day's temperature was 0° centigrade.
I. 75 on an economics test.
J. John earns $35 an hour.

1.4.2   List the classification of the data collected on the 180 Mecca College students.

1.5 SUMMARY EXERCISES

For each of the following paragraphs comment on: (a) degree of understanding transmitted to you, (b) problems in the paragraph for you; (c) the kind of data used, (d) any limitations these data cause, and (e) statistical problems you need to consider before agreeing or disagreeing with the author.

1.5.1 A study of workers finds 30% of them heart attack risks, *N.Y. Times,* March 22, 1970.*

A study of 10,000 workers in the Chicago area has found that 3000 of them can be considered high risk candidates for heart attacks.

The finding came out of a two-year pilot study aimed at identifying and reducing the high cost of cardiovascular disease to business and industry.

The American Heart Association, which released details here, said about 30 per cent of the manufacturing and office workers screened had two or more of the risk factors associated with heart attacks.

"This means," the association said, "that their chances of suffering an attack are more than double those of persons with no risk factors."

Of those in the high risk bracket, 10 per cent of 1000 persons had three or more risk factors, indicating that their risk of suffering a heart attack is as much as 10 times normal.

"If the Chicago findings are typical of the entire nation," the association said, "the national work force of about 80 million would include about 24 million in the high risk bracket, and about 8 million in the very high risk bracket."

1.5.2 "Boyfriend's attitude disturbing to her," Dr. Joyce Brothers, *Durham Sun,* July 3, 1970.

Dear Dr. Brothers: My boyfriend and I have been seeing each other constantly for three months. When we first started dating, it was he who seemed most involved. When he told me that he was in love with me, I still wasn't sure what I felt for him.

But now that I know that I'm in love too it bothers me that he doesn't seem to show his feelings very much. Sometimes he seems more interested in playing tennis than being with me and every once in a while he forgets to call me when he promised because he gets absorbed in a book.—R.S.

Dear Miss S.: From the time that a couple decides that they are in love to the conclusion of their courtship, whether in marriage or the break-up of the relationship, there is bound to be a certain amount of testing of each other's depth of feeling and commitment. Obviously, declarations of love alone are not enough. Behavior must appear to express the love that each feels.

However, a man and a woman may differ in what behavior they consider to be indicative of love. A recent survey of men and women students at a large Midwestern university attempted to determine how men and women felt when they were in love.

One finding was that male students tended to recognize and accept feelings of

*©1970 by The New York Times Company. Reprinted by permission.

love earlier in the relationship than did female students. Approximately 40 percent of the males, in comparison with 29 percent of the females, reported realizing the existence of love feelings in the beginning stages of the relationship. Approximately 43 percent of the women and 30 percent of the men failed to recognize feelings of love until after 20 or more dates.

On the other hand, the survey found that women, once they acknowledged love feelings, were more likely than men to idealize and romanticize the relationship. Both sexes reported feelings of general well-being, difficulty in concentrating, exuberance and giddiness as a result of their being in love. Women, however, reported experiencing these feelings with a greater intensity than men did.

The authors of the study hypothesized that this difference in the intensity of romantic feelings and the tendency of women to idealize the men they were in love with might be due to the greater social emphasis put upon love and marriage as significant experiences for women.

A man is not expected to let his feelings of love completely dominate his life to the detriment of other interests and commitments, while romantic preoccupation is tolerated, even encouraged in women.

Your dissatisfaction with your boyfriend's failure to behave in a suitably romantic way seems based on idealized standards of what is appropriate behavior for a person in love.

1.5.3 "State says many earn under $100—25% of workers get less than 'lower standard,'" Peter Kihss, *N.Y. Times*, March 22, 1970.*

One of every four full-time workers in private jobs in New York City and State was earning less than $100 a week last fall, according to a new State Labor Department study.

This means that these employees, for working 52 weeks, would have got less than $5200 a year, including overtime pay. The Federal Bureau of Labor Statistics has estimated that a family of four needed $6771, as of last spring, to meet costs of a "lower standard" of living in the New York–Northeastern New Jersey area.

The new state study indicated half the workers in industries covered by Federal and state minimum wage laws earned less than $2.92 an hour in straight-time pay in the city last fall. The comparable figure statewide was $2.83.

The state analysis became available at a time when rising living costs—the Consumer Price Index for this area has gone up 7.6 percent since February, 1969—have spurred increasing labor unrest, including a letter carriers' strike. The mailmen's starting pay is $6176 a year—$2.97 an hour.

The state study indicates that one of every 10 full-time workers in the city was earning less than $80 a week last fall. A four-person family on welfare rolls in the city receives a basic $208 grant a month plus an average of $100 a month for rent, a total working out to $77 a week.

Governor Rockefeller has been asking the Legislature to increase the state's hourly minimum wage.

The Governor's proposal calls for a state minimum of $1.85 an hour effective next July 1, compared with a current state and Federal minimum of $1.60, which has been in effect since Feb. 1, 1968. On a 40-hour week, the Governor's plan would mean a minimum wage of $75 a week, up from $64.

Earnings as of October. The new study by the State Labor Department's Division of Research and Statistics, whose director is Charles A. Pearce, offers estimates of employee earnings in private industry as of last October, both for workers covered by minimum wage laws and exempt workers.

For full-time workers in New York City, defined as those working 30 or more hours a week, gross weekly earnings were estimated as follows:

Earnings	No. Workers	Per Cent
Under $50	3,333	0.1
$50 and under $60	41,492	1.6
$60 and under $70	90,727	3.5
$70 and under $80	138,253	5.4
$80 and under $90	177,081	6.9
$90 and under $100	192,359	7.5
$100 and under $125	420,741	16.3
$125 and under $150	357,306	13.8
$150 and over	1,159,208	44.9
Total	2,580,500	100.0

Counting 384,100 part-time workers as well, the city had 2,964,600 employees in private industry. Of the over-all total, 985,253, or 33.3 percent, earned under $100 a week.

Of the city's full-time workers, 643,245, or 24 percent, earned less than $100. The median earnings—that is, half the workers had more and half less—were estimated as $129.03 over all, full-time workers getting $140.76 and part-timers $53.27.

The study included estimates of median straight-time hourly pay for workers in each of 87 industries under the minimum wage laws. Within New York City, those under $2.50 an hour included:

Industry	Workers (in Thousands)	Hourly Median
Rubber & miscellaneous plastic products mfg	11.2	$2.11
Leather & leather products mfg	30.6	2.12
Variety stores	10.0	1.81
Other general merchandise stores (excl. department stores)	11.3	2.22
Food stores	61.9	2.22
Apparel & accessories stores	55.5	2.19
Drugstores	6.3	2.28
Eating and drinking places	116.2	2.18
Residential buildings	40.4	2.40
Laundries, dry cleaners	24.8	2.01
Beauty shops	13.2	2.17
Barbershops	4.1	2.33
Shoe repair, hat cleaning	1.0	1.93
Miscel. personal services	2.1	2.47
Temporary help agencies	21.4	2.26
Motion picture theaters	6.0	1.86
Dance halls, studios, schools	1.0	2.46
Bowling and billiards	1.7	1.96
Convalescent and rest homes	9.7	2.47

1.5.4   A *Newsweek* poll: Mr. Nixon holds up, *Newsweek*, May 25, 1970, p. 30.*

Even after the Cambodian invasion and the killings at Kent State University, the "silent majority" appears to be alive and well in Richard Nixon's corner. A *Newsweek* Poll conducted by The Gallup Organization last week suggests that— despite the recent intense criticism of the President by college students and academic leaders and by liberal politicians and commentators—Mr. Nixon's standing with the electorate remains undamaged. The poll indicates that Americans find Mr. Nixon's conduct of the Presidency "satisfactory" by better than 2 to 1, that 50 percent favor the Cambodian operation and 39 percent oppose it, that a strikingly large majority is far more willing to blame student demonstrators than National Guardsmen for the deaths of four students at Kent State, and that Vice President Spiro Agnew's rhetoric about dissenters still enjoys the approval of a silent plurality if not a majority.

To get swift results, the survey was conducted by telephone on May 13 and 14 and covered a scientifically selected national sampling of 517 persons.**

** Telephone surveys, it should be noted, contain a slight built-in bias—about two percentage points, in this case—in favor of Republicans, since nontelephone households are necessarily omitted from the sample and these tend to be low-income and Democratic.

Although the poll gave the President majority approval of his decision to send U.S. troops into Cambodia, the favorable rating was by no means as high as some opinion experts have come to expect after dramatic strokes of U.S. military power, when Americans have a tendency to rally round the President. Following the air raids on North Vietnam that President Johnson ordered in 1965, for example, public approval (as measured by Louis Harris) soared to 83 percent. And 69 percent (polled by Oliver Quayle) favored the entry of U.S. troops into the Dominican Republic.

Women were far more dovish than men on the Cambodian issue. They opposed the President's action, 49 to 37 percent, while men supported it, 63 to 30. Women also tended to be distinctly less enthusiastic about the Vice President's speeches on dissent: in a near even split (37 to 35 percent), they approved the Veep's line, whereas men applauded him by a margin of more than 2 to 1. Young people, too, were predictably more skeptical of the Administration than their elders, but even in the 21–34 age bracket, 55 percent gave the President a favorable rating and 49 percent approved of Cambodia. And if youth was by no means arrayed entirely on the left, neither were blue-collar workers all to the right: those without a high school education came down hard against Mr. Nixon's Cambodian policy. A hefty 56 percent opposed it, and only 26 percent approved.

The question on the Kent State killings produced an unusually high number of 'no opinions,' suggesting that the no-opinion column might harbor some people with qualms about the guard's behavior who were reluctant to say so outright. It also seems likely that some of those polled were suspending judgment about who was most to blame until the conflicting accounts of the shooting could be cleared up. But even if all those with no opinion were added to those who pinned major responsibility on the National Guard, a surprisingly strong majority of each group—by age, sex, education and political party—put the main blame on the protesters.

Nixon as President		U.S. Troops in Cambodia		Who's to Blame at Kent		Agnew's Stand	
How satisfied are you with the way Richard Nixon is handling his job as President?[a]		Do you approve or disapprove of President Nixon's decision to send American troops to Cambodia?		Who do you think was primarily responsible for the deaths of four students at Kent State University?		Do you approve or disapprove of Agnew's stand on dissenters and student protesters?	
Very satisfied	30%	Approve	50%	The National Guard	11%	Approve	46%
Fairly satisfied	35%	Disapprove	39%	Demonstrating students	58%	Disapprove	30%
Not too satisfied	18%	No opinion	11%	No opinion	31%	No opinion	24%
Not at all satisfied	13%						

[a] Undecided not shown

1.5.5 "Lottery not random," Letters to the Editor of *The Times, N.Y. Times,* December 1969.*

To the Editor:

Inspection of the draft lottery results clearly shows a systematically increasing number of men being drafted as their birthdate falls later in the year. The odds against this trend resulting from random selection are over 100,000 to one.

For example, twice as many men with December birthdates will be drafted, compared with those having January birthdates. This can be easily seen by plotting the average monthly draft number from January through December. The plot gives a nearly linear decrease in average age draft number (increasing draft risk) with date of birth.

It is as if the capsules containing the birthdates were placed in the glass bowl in monthly order with January on the bottom and December on the top and then mixed or stirred too little for a random mix to be obtained.

The monthly average draft numbers from January to December are approximately: 201, 203, 226, 204, 208, 196, 182, 173, 157, 182, 140 and 122. Note that the first six months all have averages above the over-all average of 183.5, and the last six months averages are all below the over-all average.

The coefficient of linear correlation between the order number of the lottery drawing and the order of the birthdate from January is -0.222, with a standard deviation of 0.052. If the drawings were random this coefficient would be very near zero. The chance of the coefficient being this far from zero is less than one in 100,000.

Men born in November and December with draft numbers below 184 should be given a new deal by having their 47 birthdates redrawn from a new lottery which would give them order numbers to be multiplied by 366/47 and then interlaced with the remaining present numbers. The October numbers show a statistical fluctuation toward fairness.

Without this or a similar remedy these men will be subjected to an unfavorably biased treatment in opposition to the intent and spirit of the lottery.

Fred T. Haddock
Professor of Astronomy

University of Michigan
Ann Arbor, Mich., Dec. 5, 1969

1.5.6 "18–20 Group Lags on Registering to Vote," Edward C. Burks, *N.Y. Times,* August 13, 1972.**

The city has 360,000 residents who are eligible to vote for the first time as a result of the lowered voting age, but they are far more apathetic than their elders about being registered.

According to the Board of Elections, during 1971—the first year that people aged 18, 19 and 20 could register—only a third of them did so.

On the other hand, nearly 58 percent of all New Yorkers over the age of 21 are registered.

Thus there has been a rather lackadaisical response by youth to the lowering of the voting age from 21 to 18, which came into effect with the ratification of a Constitutional amendment last year. The response could have profound implications for the Democratic party, which has been counting on new youthful voters to help supply a margin of victory this fall.

The new voters were, as expected, overwhelmingly Democratic in their party preference. The board's tabulations showed 64 percent of them signing up as Democrats, roughly the same percentage as their elders. In addition, 11 percent were recorded as Republicans, 7 percent as Liberals and 2.6 percent as Conservatives. More than 15 percent did not list a party affiliation.

Analyzing 1970 census figures, the Community Council of Greater New York has determined that half of the city's future young voters—those now younger than 18—live in poverty areas.

Each year, approximately 120,000 New Yorkers turn 18. Thus, at any given time in recent years the "pool" of young people aged 18, 19 or 20 has been about 360,000.

Using projections from the 1970 census figures, The New York Times has prepared maps, which show heavy concentrations of people in the new voting-age bracket in black and Puerto Rican neighborhoods designated as official poverty areas.

The maps are divided into community planning districts, which generally coincide with one or several recognized communities. The district making up Bedford-Stuyvesant in Brooklyn, the city's largest black community, has the highest number of 18-to-20-year-olds, according to census projections—a total of 12,193.

There is no exact count of those aged 18 to 20 living in each of these districts today. However, everyone was tabulated by age at the time of the 1970 census, and demographers believe that a reasonable estimate can be made of the present situation by using 1970 totals of young people who at that time were 16, 17, and 18—and who are now 18, 19 and 20.

The projections show major clusters of the 18-to-20 group in these other districts: Williamsburg, South Brooklyn, Crown Heights and East New York in Brooklyn, Morrisania in the Bronx, South Jamaica, Flushing, Hollis, St. Albans, Queens Village and nearby areas in Queens.

Alexander Bassett, administrative manager of the Board of Elections, said that board personnel found great apathy in the poverty areas when trying to sign up new voters.

Referring to last summer's major effort, when 25 vans were sent into neighborhoods, Mr. Bassett said it was an expensive "flop" because the program cost $250,000 and only some 40,000 people were registered.

Up to last Monday, the total number of registrants all over the city this year was slightly less than 160,000. That figure includes those registering for the first time as well as those signing up again after moving or allowing their registration to lapse.

Unless there is a considerable spurt in new registrations prior to the final cutoff in October (after four days of neighborhood registration Oct. 5, 6, 7 and 10), the number of New Yorkers eligible to vote for President will not be very much greater than in 1968.

New registration is offset because about 10 percent of those on the list are purged each year, according to Mr. Bassett.

Generally, a large proportion of the new registrants are those who have just reached voting age.

In 1971, when the Board of Elections kept a rather complete score, it found that 29 percent—or 127,440 of the 440,000 registrants—were those in the new voting-age bracket.

They were distributed by borough as follows: Manhattan, 14,885; Bronx, 21,881; Brooklyn, 44,635; Queens, 40,812; and Richmond, 5227.

This year only two of the boroughs—Queens and Richmond—continued to keep a count of the number of young people registering. Queens up to Aug. 7 had 11,912; and Richmond, 2059.

At the same time, however, Queens had more than 30,000 and Richmond more than 5000 young people becoming 18 this year and able to register. The figures show far less than half of the new group of 18-year-olds are registering.

1.5.7 "Do poor people have more friends?", Shirley Sloan Fader, *Family Weekly* of the *Sunday Record*, Troy, New York, July 21, 1974.*

If you enjoy watching "The Waltons," you're involved with the current nostalgia fad that shows the Depression years as a time of great human warmth. Though most nostalgia about any era is unrealistic, the family-friend warmth idea connected with poverty may actually be true. A study of 4500 modern families' leisure habits shows that to this day poorer people regularly out-socialize more prosperous families. The poor very frequently drop in on their neighbors, relatives and friends for an informal visit of helping out, TV watching or just sitting around; and they definitely keep closer overall visiting contact with their relatives than does the average prosperous person. People with good incomes drift away from spending time with friends and kinfolk toward leisure activities that must be bought and paid for—bowling, golf, trips, restaurants, movies, etc. It's true that purchased entertainment is often interesting and fun. But the old idea of spending time with friends still satisfies a basic human need.

1.5.8 "The dream of many, the realization of few," Barbara Hyland, *The Stanford Daily,* © 1968.

*Reprinted by permission of the Family Weekly Magazine.
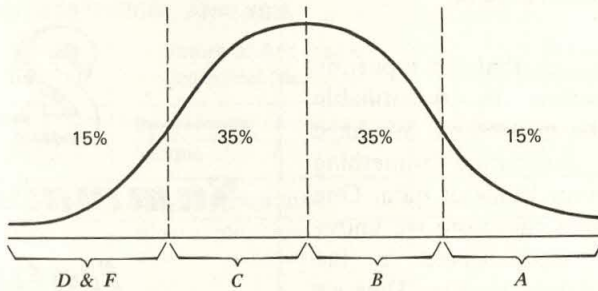
"The Perfect Crime" was almost committed at Stanford this Fall. In true Robin Hood fashion, with the cause of studenthood as their ideal, the heroes tried to finagle the grading in Psychology 1 so that no one failed.

Their racket consisted of taking midterms and quizzes under fictitious names. By doing poorly, they could bring down the curve. They even handed in computer cards and blue cards so that the fictitious names would be placed on all of the computerized class lists.

"If we get enough people to do it, no one can fail," enthusiastically explained one unidentified member of the class. Only fake people would get F's.

However, they forgot the invincible Registrar's Office, which makes up class lists according to the cards each student hands in with his student number. These lists are then sent to each department.

Acting on a hot tip-off, J. Merrill Carlsmith, one of the instructors of Psych 1, compared the Registrar's list with the list of people who had taken tests. Even if he had not checked over the lists now, he said the discrepancies would definitely have been discovered when the final marks were recorded.

Carlsmith said that the fake grades were included in calculating the mean for the midterm and some quizzes. He said, however, that there were "so few, it isn't having effect at all." Despite the grade conspirators optimistic force of 20 cohorts, their force was small. Only three fictitious people took the midterm. They got two F's and one D. Five nonexistent people took the recent quiz. One, who took both the midterm and the quiz, wrote Carlsmith a note asking if he could switch to Pass-Fail grading.

Since Psych 1 has an enrollment of approximately 300, they barely lowered the mean. In the case of the midterm, the true mean should have been about two hundredths of a point higher.

If there had been a significant number taking phony tests, thus altering the mean, Carlsmith explained that he would have just "pulled the cards and rerun them" through the computer. He remarked that it would be "an awful nuisance" if it had been done in a class that didn't have computerized grading.

So students took another loss in their struggle to outfool the authorities. But the dream still lives...

2.1 INTRODUCTION

We have seen that the reporting of information in quantifiable terms can be misleading. We have begun to understand something about different kinds of data. One might say that the more we know, the more sophisticated is the measuring device we use. Thus we can look forward to a kind of progression in our use of statistical methods of analysis (as shown opposite).

The goal of this chapter is to present some graphical techniques for reducing a mass of data down to a form that lets us see the forest as well as the trees. These techniques include various tables, charts, and line graphs. Some devices are best for one type of data, others for other types. We will then be led into Chapter 3, where we can add numerical techniques to the graphical ones.

Before we begin to discuss any techniques, we had better decide just what we hope to accomplish by any analysis we undertake. The objective of statistical analysis is to answer very well-defined special questions. Let us use the data we have collected on the $n = 180$ upper-class students at the community college and decide on some specific questions that might be addressed to the data:

2

Summarizing Data Graphically

2

1. What proportion of the sample is female?
2. What is the average commuting distance of the students in the sample?

PROGRESSION OF STATISTICAL ANALYSIS

Type of Data		Nature of Applicable Statistical Methods	Number of Statistical Tools
Discrete	Nominal	Very simple	Few
	Ordinal	Simple	A few more
Continuous	Interval	More sophisticated	Many more
	Ratio	Most sophisticated	Most possibilities

3. What is the longest distance a student commutes? What is the shortest distance?
4. Can we obtain an overall picture of student commuting distances?
5. Is there any difference between female commuting distance and male commuting distance?
6. Is there any relationship between a student's G.P.A. and the distance he commutes?
7. What is the political preference profile for these 180 students?
8. Is there any difference in the political preferences of men and women in the sample?
9. What is the position of the 180 students on the question of legalizing marijuana?
10. Is the overall position on legalizing marijuana affected by the sex of students or by student G.P.A.?
11. What is the G.P.A. profile of these students?
12. For each of the above questions, how can we generalize the results to the population of all upper-class students in the community-college system? Under what restrictions would we be willing to make statements about all the upper-class students?

2.2 TABULAR SUMMARIES

If we are to answer the questions listed above, we will have to take the collection of data given in Table 1.2.1 and tabulate them. We call the original observations *raw* data. They are in original form, assembled but unorganized, nutritious but uncooked. As soon as we start to tabulate information, these raw data will take on a new form; something of shape will be gained, something of detail will be lost in the transition.

For example, in order to answer the first objective question: "What proportion of the sample is female?", we will have to count up the number of females in the sample. Let us record this information in Table 2.2.1.

TITLE —————————→ **TABLE 2.2.1** **180 Community College Upper-class Students, by Sex, Mecca Community College, 1974**

Column **HEADINGS** ———→

Row **HEADINGS**

CELL

Sex	Frequency or Number	Proportion
Female	76	$\frac{76}{180} = 0.42$
Male	104	$\frac{104}{180} = 0.58$
Total	180	1.00

SOURCE —————————→ Source: Table 1.2.1

Table 2.2.1 illustrates the following characteristics of a good table:

1. *Simplicity.* The simpler the table, the better. A good rule to use is "Only the specific data needed to answer one particular question should be in any one table."
2. *Title.* Every table should have a complete title. This title should identify the table's contents by *what, where,* and *when.*
3. *Headings.* Every row and every column of the table should be labeled with a short, clear, and concise heading. If extensive detail of identification is required, use brief label plus footnote.
4. *Cells.* Every table consists of a number of subunits called "cells." A cell is the intersection of one row and one column. The table thus is constructed in such a way that any observation (piece of raw data) can be placed in one and only one cell of the table.
5. *Source.* If the data are being summarized from another source, the complete source must be put as a footnote to the table.

The following table was constructed for the tabulation (summary) of *nominal* data. *Ordinal* data can be summarized in similar fashion.

TABLE 2.2.2 Full-Time Enrollment of Persons 14–34 Years
Old in Two-Year Colleges, By Population
Density of Place of Residence, United
States, October 1972
(Numbers in Thousands; Civilian Noninstitu-
tional Population)

Residence Density	Male	Female	Total
Metropolitan areas inside central cities	219	149	368
Metropolitan areas outside central cities	356	208	564
Nonmetropolitan areas	195	127	322
Total	770	484	1254

Source: Undergraduate enrollment in 2-year and 4-year colleges: October 1972,
U.S. Bureau of the Census, *Current Population Reports*, Series P-20, No. 257, U.S.
Govt. Printing Office, Washington, D.C. (1973), p. 15.

When one makes tables for *interval-* or *ratio-scale data*, another dimension of complexity is added. It is not always clear how one should organize the row headings. For example, let us consider summarizing in tabular form the commuting distance information on the 180 students of Table 1.2.1. The following steps are suggested. After finishing them all, we will retrace the steps, perhaps eliminate some, and take shortcuts on others.

STEP 1. Let us first reorder the commuting distances from the shortest distance to the longest distance (Table 2.2.3). We note that the shortest distance is 0.5 mile and the longest distance is 56.5 miles. We need to make up categories of distance so that we have a better understanding of the commuting distances.

TABLE 2.2.3 Commuting Distances Ordered From Small to Large

0.5	2.5	4	5.5	7.5	11	15	20.5	27.5
0.5	2.5	4	5.5	7.5	11.5	15	21.5	27.5
0.5	2.5	4	6	7.5	11.5	15	21.5	28.5
0.5	3	4	6	7.5	11.5	15	22.5	28.5
0.5	3	4	6	8	12	15	22.5	28.5
0.5	3	4	6	8	12	15.5	22.5	28.5
1	3	4	6	8	12	15.5	23.5	29.5
1	3	4	6	8	12	16.5	23.5	29.5
1	3	4	6	8.5	12	17	24	29.5
1	3	4	6.5	9	12.5	17	24	32
1	3	4.5	6.5	9	12.5	17	24	32
1	3.5	4.5	6.5	9.5	12.5	17.5	24	32
1	3.5	4.5	7	9.5	12.5	19	25	35
1	3.5	5	7	9.5	13	19.5	25	35
1.5	3.5	5	7	10	13	19.5	25.5	36.5
1.5	3.5	5	7	10	13	19.5	25.5	36.5
2	3.5	5	7	10	13.5	20	25.5	36.5
2	3.5	5	7	10.5	13.5	20	26.5	46.5
2	3.5	5	7	10.5	13.5	20.5	26.5	56.5
2	3.5	5.5	7.5	10.5	14.5	20.5	27.5	56.5

STEP 2. The next step is to decide on a method of grouping the data into subsets representing intervals on the distance scale. You must be careful in selecting the boundaries of the intervals so that every distance in the data set falls into one and only one interval. In most cases this is easily done by noting the measuring units in which the data are taken and then using the next more refined unit of measurement for boundaries. For example, distance here has been measured to the nearest one half mile, so the intervals will be constructed using one fourth mile in boundaries.

One further point needs to be considered: how many intervals should be selected? There is no simple answer to this question. For example, if an honor society had a rule that only students with at least a 3.00 G.P.A. were eligible to join, it would divide the data on G.P.A. in Table 1.2.1 into two intervals, 0–2.99 and 3.00–4.00. If the State were interested in stimulating attendance at the community college by paying some transportation costs according to the following schedule, the data would be intervalized in accordance with that schedule. In this way, some idea of the ultimate cost could be developed.

The chief criterion for selecting the number of intervals should be "How do the data have to be organized so that I can get some idea of the answer to the main question being asked about the data?" In cases where there is no basis for such a decision the usual rule is to take from 8 to 20 equally spaced intervals, with the criterion being to smooth the picture that would be obtained by plotting the data. Convenience of tallying data and reading pictured scales is taken into

account as well; thus 5–10–15–20–25 and so on is more convenient intervaliz-
ing than 3–7–11–15–19 and so on.

Miles	State contribution ($ per month)
0–4.99	10
5–14.99	30
15–29.99	60
30–44.99	90
45 and over	120

For the data on distance there are no special conditions to affect our choice
of intervals. On examining the data, we note that the shortest distance is 0.5 mile
and the longest, 56.5 miles. In order to cover all distances, let us divide 60
miles into 12 intervals of 5 miles each, starting with 0.25 miles. Then the
intervals look as follows:

0.25   5.25   10.25   15.25   20.25   25.25   30.25   35.25   40.25   45.25   50.25   55.25   60.25

The data are then tallied into the various intervals, one observation at a
time. Both the tallying and the summarizing of results are best accomplished in
tabular form:

TABLE 2.2.4 Commuting Distances (in miles) of Upper-class College Students

Distance (miles)	Tally	Frequency
0.25–5.25	JHT JHT JHT JHT JHT JHT JHT JHT JHT JHT JHT JHT IIII	59
5.25–10.25	JHT JHT JHT JHT JHT JHT JHT JHT III	38
10.25–15.25	JHT JHT JHT JHT JHT JHT III	28
15.25–20.25	JHT JHT III	13
20.25–25.25	JHT JHT JHT I	16
25.25–30.25	JHT JHT JHT	15
30.25–35.25	JHT	5
35.25–40.25	III	3
40.25–45.25		0
45.25–50.25	I	1
50.25–55.25		0
55.25–60.25	II	2
	Total	180

If we now use Table 2.2.4 in place of the original raw data, we have sacrificed some of the original information. For example, we know that 59 out of the 180 sampled students travel between 0.25 and 5.25 miles each way to campus, but we do not know anything about how the 59 distances are distributed along the interval. We have to assume that they are evenly distributed across the interval. So while tables help to clarify and summarize information, they do sacrifice some information in the process.

EXERCISES

2.2.1 Construct a table for showing how the opinion on legalizing marijuana differs with respect to sex.

2.2.2 Using the data in Table 1.2.1, fill in the following tables by tallying the numbers of students for the various cells.

Political Party Preference

Party Preference	Male	Female
Democrat		
Republican		
Other Party		
No Preference		

Opinion on Legalizing Marijuana

1 Strongly Disagree	2 Mildly Disagree	3 No Opinion	4 Mildly Agree	5 Strongly Agree

2.2.3 Based on what you find in the above three tables, discuss whether the data in these tables support or do not support the following statements:
   a. In general, this sample of 180 students shows that young people would like to see marijuana legalized.
   b. While men prefer the legalization of marijuana, women do not.
   c. Men are more committed to political parties than women are.
   d. This sample has too many men in it; thus any ideas about generalizing should not be allowed.
   e. This sample of 180 students shows some interesting opinions, but I don't believe that all of the students would be similar to these students.

2.2.4 The following questionnaire was distributed to Rensselaer Polytechnic students in the spring of 1973 by the (student) Union Programs Activity Committee (UPAC).

> How to get into a movie for nothing. UPAC is publishing the following questionnaire in an attempt to find out what the students prefer in entertainment. Future programming will reflect these opinions.
> Any student bringing a completed questionnaire to Friday night's showing of "Beneath the Planet of the Apes" will receive free admission for himself

and his date. Other students may return the completed questionnaires to the boxes at the Union Information desk and elsewhere on campus.

Please circle the correct answer:

1) What school are you from?
   RPI
   Sage
   Other_____

2) If from RPI, where do you live?
   Dorms
   Fraternity
   Off-campus

3) Are you a(n)?
   undergraduate
   graduate

4) Are you?
   male
   female

5) Are you?
   single
   married

6) If married, how many children over 2 years of age do you have? (please fill in number)_____

7) Did you know what UPAC was before reading this article?
   yes
   no

8) Do you attend UPAC movies?
   always
   frequently
   infrequently
   never

9) Did you, in general, like UPAC's movie selection?
   yes
   no
   no opinion

10) Do you attend UPAC speakers?
   always
   frequently
   infrequently
   never

11) Did you, in general, like UPAC's speaker selection?
yes
no
no opinion

12) Do you attend Beer & Flicks?
always
frequently
infrequently
never

13) Do you attend Foreign Films?
always
frequently
infrequently
never

14) Do you attend Coffee Houses?
always
frequently
infrequently
never

15) Do you go to bands in the Rathskeller?
always
frequently
infrequently
never

16) Do you attend UPAC Cultural Events?
always
frequently
infrequently
never

17) Which did you attend: (Circle those that apply)
J. Geils
Isaac Hayes
Chicago
Chuck Mangione
Gary Burton

18) Have you, in general, been satisfied with UPAC's level of programming:
yes
no
no opinion

19) If you wanted to suggest some type of program for UPAC, would you know where to go or whom to see?
yes
no

Columns

Student Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
284	1	3	2	1	1	0	1	3	1	2	1	3	4	4	2	2	0	0	0	1	1	1	1	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0
285	1	3	1	1	1	0	1	3	1	3	1	4	4	4	2	3	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
286	1	2	2	1	1	0	1	4	3	4	3	4	4	4	4	4	0	0	0	0	0	3	2	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
287	1	3	1	2	1	0	2	4	3	4	3	4	4	4	4	4	0	0	0	0	0	3	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
288	1	1	1	1	1	0	1	2	1	4	3	3	4	4	4	4	0	0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	1
289	1	2	1	1	1	0	2	2	1	3	3	3	4	4	4	4	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1
290	1	3	1	1	1	0	1	3	3	4	3	3	4	2	3	3	1	1	1	0	0	0	1	0	1	1	0	0	1	1	1	0	1	0	1	1	0	0	1
291	1	1	2	1	1	0	1	2	1	4	1	3	4	3	2	2	1	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
292	1	3	1	1	1	0	1	4	1	4	3	4	4	2	2	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
293	1	1	2	2	1	0	1	3	1	4	3	3	4	4	3	3	0	0	1	0	0	3	2	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1
294	1	2	1	1	1	0	1	4	3	3	3	4	3	2	2	2	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
295	1	1	2	2	1	0	2	3	1	4	1	4	4	3	2	2	1	1	1	0	1	3	2	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	1
296	1	3	1	1	1	0	1	4	3	4	3	3	4	3	3	3	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0	1
297	1	1	2	1	1	0	1	4	1	4	1	1	4	4	4	4	0	0	1	0	1	2	2	0	0	0	0	0	1	0	1	1	0	0	1	1	0	0	0
298	1	3	1	1	1	0	1	4	1	4	3	3	4	2	3	3	0	0	1	0	0	3	1	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0
299	1	3	1	1	1	0	1	4	3	3	3	4	3	3	4	4	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1
300	1	1	1	2	1	0	2	3	3	4	1	1	4	2	4	4	0	1	1	0	0	2	1	0	0	1	0	1	1	0	1	1	0	0	1	1	0	0	1
301	1	3	1	1	1	0	1	2	1	3	3	3	4	4	3	3	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0
302	1	1	1	1	1	0	1	3	1	4	1	4	4	3	4	4	0	0	0	0	1	3	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
303	1	3	1	1	1	0	1	2	3	3	3	4	3	4	3	4	0	0	1	0	0	2	1	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0
304	1	1	1	2	1	0	2	4	3	4	3	3	4	4	4	4	1	0	1	0	1	3	1	0	0	1	0	0	0	0	1	1	0	0	0	1	0	1	1
305	1	1	1	1	1	0	1	2	3	3	1	4	4	2	3	3	0	0	1	0	0	2	2	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	1
306	1	1	1	1	1	0	1	2	3	3	3	3	4	4	4	3	0	0	0	0	0	3	2	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
307	1	1	1	1	1	0	1	2	1	4	1	4	4	4	4	4	0	0	1	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
308	1	1	1	1	1	0	1	4	1	3	3	3	4	2	3	4	0	0	0	0	0	3	1	0	1	1	0	0	0	1	1	1	1	0	0	1	0	0	0
309	1	1	2	1	1	0	1	2	2	4	2	3	4	4	3	4	0	0	1	0	0	2	1	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0
310	1	2	1	2	1	0	1	3	1	3	1	3	4	2	3	3	0	0	0	0	0	3	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0
311	1	1	1	1	1	0	1	2	1	4	2	4	3	4	2	4	1	0	0	1	0	3	1	0	1	1	0	1	0	0	1	1	1	0	0	1	0	0	0
312	1	1	1	1	1	0	1	3	1	3	3	2	4	4	2	3	0	0	1	0	0	1	1	0	1	1	0	0	0	1	0	1	0	0	0	0	0	1	0
313	1	1	1	1	1	0	1	2	1	3	1	3	4	4	2	4	0	0	1	0	0	1	1	0	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0
314	1	1	1	1	1	0	1	2	1	4	2	3	4	4	3	3	1	0	1	0	0	3	2	0	0	1	0	0	0	0	1	1	1	0	0	1	0	0	0
315	1	2	1	1	1	0	1	3	1	3	3	3	3	4	3	4	1	0	1	0	0	3	1	0	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0
316	1	1	1	1	1	0	1	2	1	4	1	2	4	4	3	3	0	0	0	0	0	1	1	0	1	1	0	0	0	1	0	1	1	0	0	0	0	1	0
317	1	1	1	1	1	0	1	3	1	4	3	3	4	4	3	4	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	1	1
318	1	1	1	1	1	0	1	2	1	3	3	2	3	3	2	3	1	1	1	0	0	3	2	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
319	1	2	1	1	1	0	1	3	2	3	1	3	3	4	3	4	1	1	1	0	0	1	1	0	1	1	0	0	0	1	1	1	1	0	0	1	0	1	1
320	1	1	1	1	1	0	1	2	1	3	1	3	4	4	3	4	0	0	0	0	0	3	1	0	0	0	0	0	0	1	1	1	0	0	0	1	0	1	0
321	1	1	2	1	1	0	1	2	2	2	3	3	3	4	3	4	0	0	1	0	0	1	1	0	1	1	0	0	0	1	1	1	1	0	0	0	0	1	1
322	1	1	1	1	1	0	1	3	1	2	3	2	3	4	2	3	1	1	1	0	0	1	2	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
323	1	2	1	1	1	0	1	2	1	4	3	3	4	3	3	4	0	0	0	0	0	1	2	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	1
324	1	1	1	1	1	0	1	2	2	4	1	3	4	4	2	3	0	0	0	0	0	1	1	0	0	1	0	0	0	1	1	1	0	0	0	1	1	0	1
325	1	2	2	1	1	0	1	2	1	2	1	2	4	3	4	4	0	1	1	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0
326	1	1	1	1	1	0	1	2	1	4	3	3	3	4	3	3	0	0	0	0	0	1	1	0	0	1	1	1	1	0	1	1	1	0	0	1	0	1	1
327	1	2	1	1	1	0	1	2	1	2	3	2	3	3	2	3	1	1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1
328	1	2	1	1	1	0	1	2	1	3	3	3	4	4	3	3	0	1	0	0	0	1	2	0	0	1	1	1	1	0	0	0	1	0	0	1	0	1	0

33

20) Circle those areas in which you would like to see more programming.
   Coffee Houses
   Beer and Flicks
   Popular Films
   Foreign Films
   Speakers
   Cultural Events
   Bands
   Big Concerts

21) How do you find out about UPAC's events?
   Poly article
   Poly Calendar Events
   Other newspapers, radio or T.V.
   UPAC Calendar of Events
   UPAC Bulletin Board
   UPAC Posters
   Word of mouth

The data were collected from a total of 550 students. The table on p. 33 is an excerpt from a computer printout of the data for students Number 284–328 inclusive. Examine the data sheet. Without the coding scheme, can you make any sense out of the data?

2.2.5 The following coding scheme was used for the computer:

Question	Category	Code	Column on Data Sheet
1	RPI	1	
	Sage	2	1
	Other	3	
2	Dorms	1	
	Fraternity	2	2
	Off campus	3	
3	Undergraduate	1	
	Graduate	2	3
4	Male	1	
	Female	2	4

Question	Category	Code	Column on Data Sheet
5	Single	1	5
	Married	2	
6	Number of children		6
7	Yes	1	7
	No	0	
8	Always	1	
	Frequently	2	8
	Infrequently	3	
	Never	4	
9	Yes	1	
	No	2	9
	No opinion	3	
10	coded like 8	recorded in column 10	
11	coded like 9	recorded in column 11	
12	coded like 8	recorded in column 12	
13	coded like 8	recorded in column 13	
14	coded like 8	recorded in column 14	
15	coded like 8	recorded in column 15	
16	coded like 8	recorded in column 16	

Question 17	Code	Column
J. Geils	1 attended	17
	0 did not attend	
Isaac Hayes	1 attended	18
	0 did not attend	
Chicago	1 attended	19
	0 did not attend	
Chuck Mangione	1 attended	20
	0 did not attend	
Gary Burton	1 attended	21
	0 did not attend	

Question 18 coded like question 9 and recorded in column 22; question 19 coded like question 7 and recorded in column 23. Questions 20 and 21 use the code 1 for yes and 0 for no, recorded in columns as follows:

Question 20	Column
Coffee houses	24
Beer & flicks	25
Popular films	26
Foreign films	27
Speakers	28
Cultural events	29
Bands	30
Big concerts	31
Question 21	
Poly (school paper) article	32
Poly calendar of events	33
Other newspapers	34
Radio or T.V.	35
UPAC calendar of events	36
UPAC bulletin board	37
UPAC posters	38
Word of mouth	39

a. Tabulate the following information:

	Frequency
Undergraduate	_____
Graduate	_____

b. Did the students like or dislike UPAC's movie selections?
c. What percentage of those who attended UPAC's movies liked the movie selection?

2.2.6 Examine magazines and newspapers and find three samples of tables of discrete data. Identify the kind of scale (nominal, ordinal, interval, ratio).

2.2.7 Using the data obtained from the survey in Exercise 2.2.5, fill in the following table:

Living Quarters	Attendance at Foreign Films				Totals
	Never	Infrequently	Frequently	Always	
Dorm Fraternity Off campus					
Totals					

Would you be willing to make some statement about the attendance at foreign films as it relates to where the students live?

2.2.8 Fill in the following table:

Attendance at UPAC Events						
		UPAC Movies				
		Never Attended 4	Infrequent 3	Frequent 2	Always 1	Totals
UPAC Cultural Events	Never Attended 4					
	Infrequent 3					
	Frequent 2					
	Always 1					
Totals						

a. How many students were frequent UPAC movie goers but infrequent UPAC cultural events goers?
b. How many students never attended movies or cultural events? What percentage of the total students does this represent?
c. What percentage of the students attended at least one UPAC movie?
d. What percentage of the students attended at least one UPAC cultural event?
e. If one defined UPAC activities as successful if attendance at both movies and cultural events was either frequent or always, would you say that UPAC was successful or not as a result of this survey? Why?

2.3 GRAPHICAL PRESENTATION

The presentation of data in tables can be further clarified to the reader by a graphical presentation of the same data. It is the authors' opinion that any interpretation or analysis of data should include a graphical presentation of the information.

A. Graphical Methods for Nominal and Ordinal Data

The most useful graphical technique for presenting nominal or ordinal scale data is the *bar chart.*

Bar Chart. A bar chart is a diagram consisting of vertical (or horizontal) bars which represent the frequency of observations in specific categories. There are several useful guidelines that should be followed in constructing a bar chart:

1. Generally the bars for categories should be separated so that there is a distinct uniform space between bars.

2. The bars must all start from the same base.

3. It is not advisable to include figures inside or on top of the bars. This creates the illusion of shortening or lengthening the bars.

4. Since a graph is drawn for the purpose of enabling the reader to gain a quick picture of the information in the data, it is advisable when dealing with nominal data to arrange the bars in ascending or descending order of magnitude.

5. If more than one color (or shading) is used, a key for colors should be prominently displayed.

6. Every bar chart should have a title, and if the data are taken from an external source, the source should be indicated at the bottom of the chart.

Let us take the table constructed on the opinion on legalizing marijuana (Exercise 2.2.2) and draw a bar chart for it.
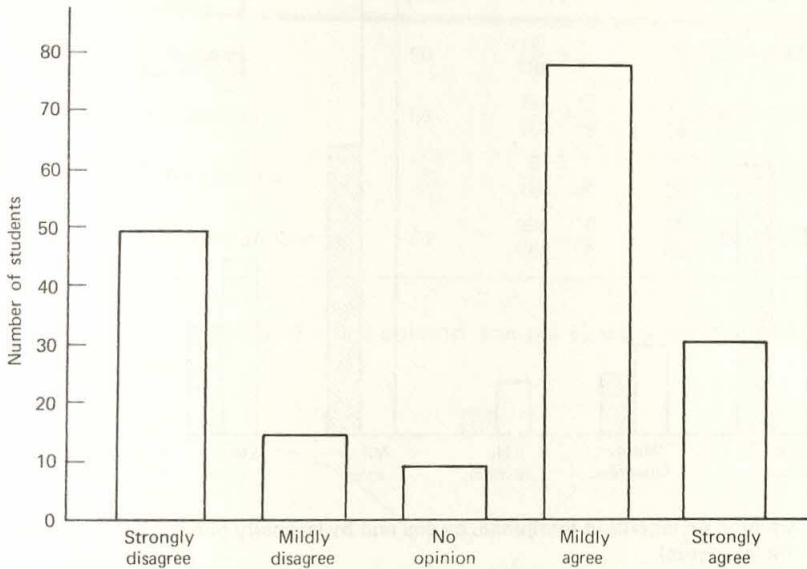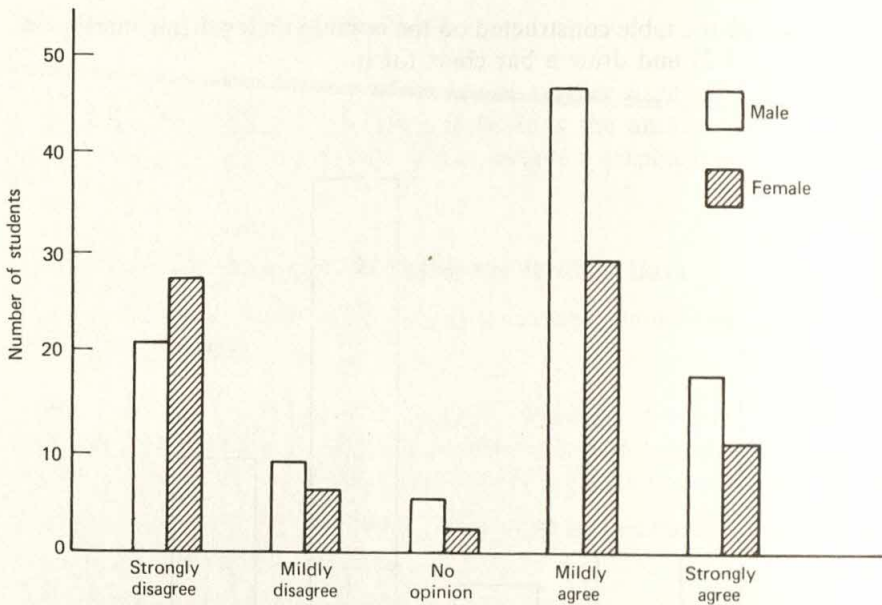

Opinion on legalizing marijuana, by intensity of opinion (sample of 180 students).

Source: Sample of 180 upper-class students, Table 1.2.1.

Note: Since the horizontal scale is a set of ordered categories, the ordering of the bars by their frequency heights is not possible.

As a further example of using bar charts, the opinion on legalizing marijuana has been further broken down by sex. This is shown in the following bar chart, where the categories of opinion are separated into sets of two bars, male and female.

Opinion on legalizing marijuana, by sex and by intensity of opinion (sample of 180 students).

Source: Sample of 180 upper-class students, Table 1.2.1.

Note: Since the horizontal scale is a set of ordered categories, the ordering of the bars by their frequency heights is not possible.

A second graphical technique useful for nominal data in particular is the *pie chart.*

Pie Chart. The pie chart is a circle in which the component percentages of the total sample are plotted by converting them to degrees. In plotting categories of nominal-scale data in this manner, there are a few guidelines which prove useful:

1. Start the division of the circle at 12:00 o'clock. The reason for this is that people tend to read a clock in a clockwise direction from 12:00 o'clock.

2. In order to put the emphasis on that category with the highest frequency (number of occurrences), the categories are usually plotted in the order of descending frequencies.

Let us use the data on political party preference to show the construction of the pie chart.

Political Party Preference

	Frequency	Fraction of Total	Degrees (Fraction × 360°)
Democrat	60	$\dfrac{60}{180} = \dfrac{1}{3}$	$\dfrac{1}{3} \times 360 = 120°$
Republican	50	$\dfrac{50}{180} = \dfrac{5}{18}$	$\dfrac{5}{18} \times 360 = 100°$
Other parties	6	$\dfrac{6}{180} = \dfrac{1}{30}$	$\dfrac{1}{30} \times 360 = 12°$
No party preference	64	$\dfrac{64}{180} = \dfrac{16}{45}$	$\dfrac{16}{45} \times 360 = 128°$

Using the degrees in the last column, the pie chart is constructed, starting at 12:00 o'clock.
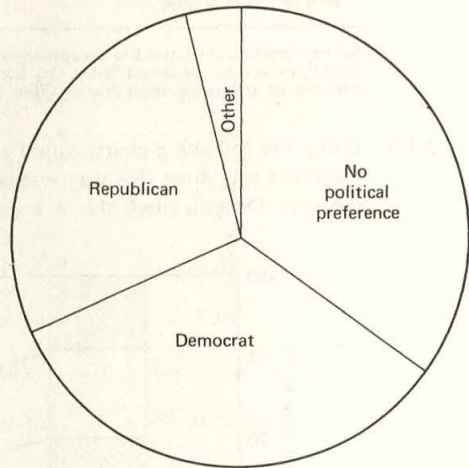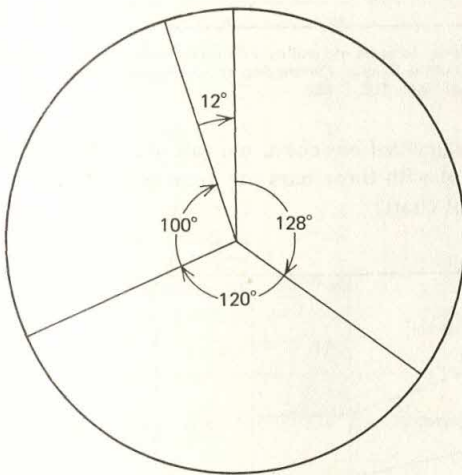
Political party preference among 180 Mecca Community College students.
*Source:* Sample of 180 upper-class students, Table 1.2.1.

EXERCISES

2.3.1 Using the data in Table 1.2.1, construct a table showing the joint breakdown of political party preference and opinion on legalizing marijuana. Prepare a graphical representation of these data.

2.3.2 Construct graphical presentation of the data from the following table:

Employment of Persons Aged 14 Years and Over as of March 1968

	Men		Women	
	Numbers in Thousands	Median Income	Numbers in Thousands	Median Income
Total	66,519	5,571	73,584	1,819
Employed	47,622	6,610	27,887	3,157
Unemployed	1,680	3,017	1,332	1,382
Armed Forces or Not in Labor Force	17,217	1,634	44,365	913

Source: Employment Status and Occupation—Persons 14 years old and over by total money income in 1967, by sex, for the United States, U.S. Bureau of the Census, Current Population Reports, Series P-60, No. 60, U.S. Government Printing Office, Washington, D.C. (1969).

2.3.3 Using the following chart, called a *segmented bar chart,* indicate why you think the chart was done this way instead of with three bars per year as indicated in the text. Do you think this is a useful chart?
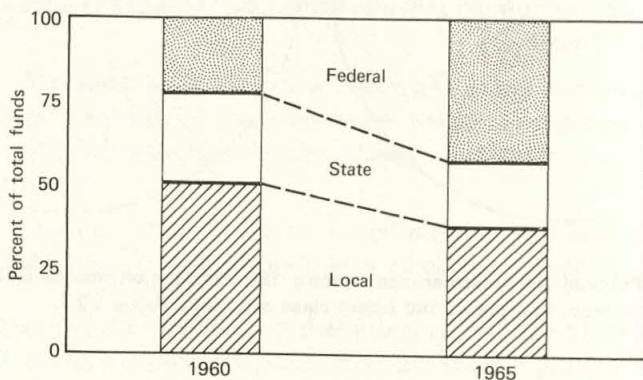
Source of funds for XYZ Health Department, 1960 and 1965.

2.3.4 Using the data shown in Exercise 2.2.4, construct: (a) bar chart showing the frequency of preference for more UPAC programming in the areas shown in question 20, and (b) pie chart showing the attendance at UPAC cultural events.

B. Graphical Methods for Interval and Ratio Scale Data

When data are recorded using a continuous scale, there are three very useful graphical methods: (a) the *histogram*, (b) the *frequency polygon*, and (c) the *cumulative frequency polygon*.

Histogram. The histogram is a bar chart with the bars *not* separated; the bases of the bars are on one continuous scale.

Let us use the commuting-distance data that we tabulated in the previous section to demonstrate the construction of a histogram. The data are repeated in Table 2.3.1 for convenience.

TABLE 2.3.1 Commuting Distances (in miles) of Upper-Class Community College Students

Interval	Frequency	Cumulative Frequency	Cumulative percentage (in decimal form)
0.25–5.25	59	59	.33
5.25–10.25	38	59 + 38 = 97	.54
10.25–15.25	28	97 + 28 = 125	.69
15.25–20.25	13	125 + 13 = 138	.77
20.25–25.25	16	138 + 16 = 154	.86
25.25–30.25	15	154 + 15 = 169	.94
30.25–35.25	5	169 + 5 = 174	.97
35.25–40.25	3	174 + 3 = 177	.98
40.25–45.25	0	177 + 0 = 177	.98
45.25–50.25	1	177 + 1 = 178	.99
50.25–55.25	0	178 + 0 = 178	.99
55.25–60.25	2	178 + 2 = 180	1.00
Total	180		

STEP 1. Plot the intervals on the horizontal axis of arithmetic line graph paper.
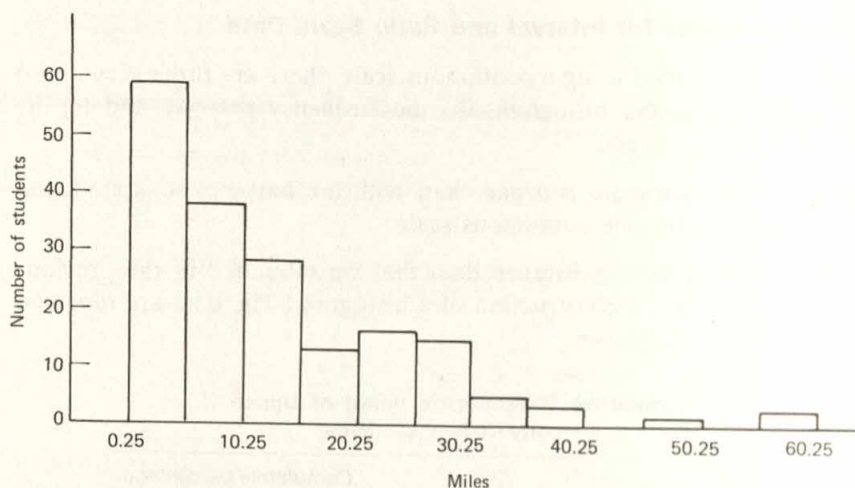
Commuting distance of 180 community college students.

STEP 2. Each bar will be 5 miles in width and have a height equal to the frequency of that interval. Thus the first bar will begin at 0.25 mile and end at 5.25 miles and have a height of 59, representing the 59 students who commute between 0.25 and 5.25 miles from their home to school. The entire histogram is shown above.

Special attention is required in constructing a histogram when the intervals are unequal in width. Adjustment of frequency plots must be made to prevent misinterpretation. The procedure is best seen in an example, as follows.

In an article there appears the following table [Auer, Eugene S., Carcinoma of the cervix uteri, *J.A.M.A.* 98 (26), 2260 (1932); the frequencies have been modified for ease of computation].

Patients Treated for Carcinoma of the Cervix, by Age
Barnard Free Skin and Cancer Hospital, St. Louis, Missouri, 1906–1926

Age	Number of Patients
22–30	16
30–35	45
35–40	79
40–55	225
55–60	63
60–70	46
70–90	12

In creating a histogram for these data, remember our assumption about the observations within each interval. In the first age interval (22–30) we have 16 patients; our assumption is that the 16 patients are spread out evenly over the 8-year interval. In the age interval 40–55, 225 patients are spread out evenly over a 15-year span. In order to plot the data to reflect this we first must choose one size interval as a standard on which to base the plotting. Usually one plots on the basis of the smallest interval cited; in our example, a 5-year age span is the choice for standard. The bar heights are adjusted to conform to the standard. For the data of the above table, the adjusted heights of the histogram are shown as follows:

| Age	Number of Patients	Number of Patients for Histogram Plot on a 5-year Basis
22–30	16	$\frac{5}{8} \times 16 = 10$
30–35	45	$\frac{5}{5} \times 45 = 45$
35–40	79	$\frac{5}{5} \times 79 = 79$
40–55	225	$\frac{5}{15} \times 225 = 75$
55–60	63	$\frac{5}{5} \times 63 = 63$
60–70	46	$\frac{5}{10} \times 46 = 23$
70–90	12	$\frac{5}{20} \times 12 = 3$

The histogram correctly plotted now comes out as follows:
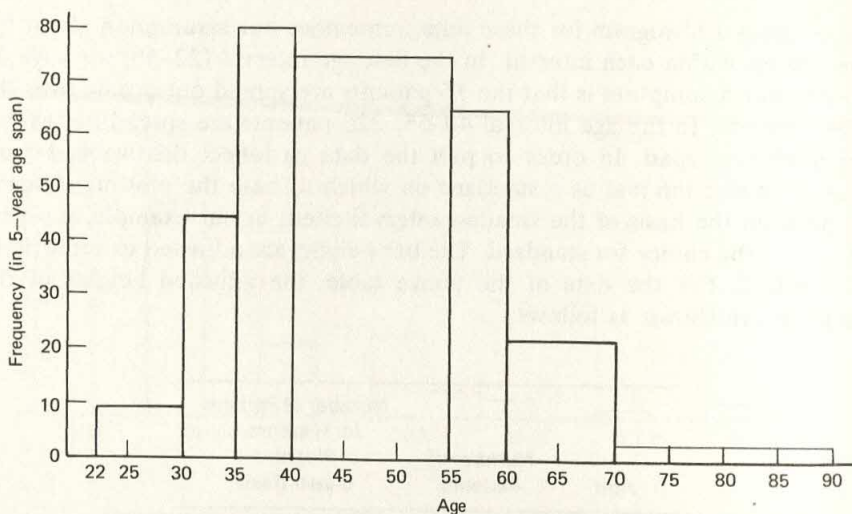
Note that the above picture is the correct one. If you plotted the observed frequencies without the adjustment, the picture would be completely misleading, since the 225 frequency would incorrectly dominate the picture. Thus, you must be very careful when drawing graphs of data which are collected in tables with unequal intervals.

The next logical kind of graph is one that replaces the histogram with a line graph called the *frequency polygon.*

Frequency Polygon. The frequency polygon is constructed by connecting the midpoints of the tops of the histogram bars with straight lines and then removing the bars.
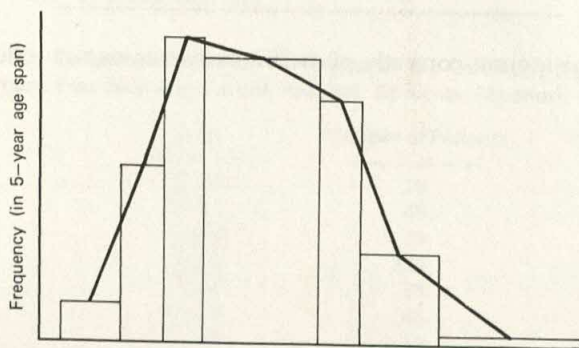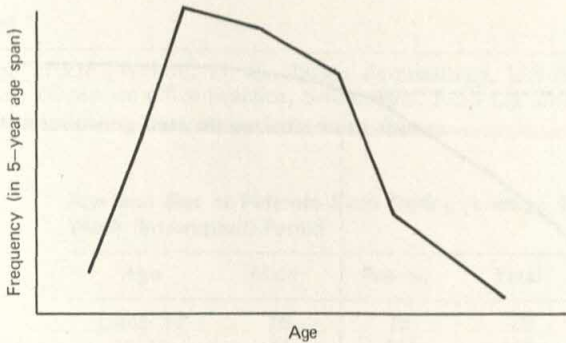
Obviously you don't have to go through the histogram routine if all you are interested in is the frequency polygon; the required points can be plotted without drawing and erasing bars.

The frequency polygon is the graphical method most often used for plotting continuous data. Please remember, however, that our polygon example has been drawn taking into account unequal intervalized data. Otherwise it too would have been as misleading as an incorrectly drawn histogram.

Cumulative Frequency Polygon. A very useful graph for many purposes is the cumulative frequency polygon. To construct this type of graph we must first calculate the cumulative frequencies. Examples are shown in the third column in Table 2.3.1.

For example, 59 students commuted less than 5.25 miles, 97 students commuted less than 10.25 miles, that is, the 59 of the 0.25–5.25 interval plus the 38 who commuted anywhere from 5.25 to 10.25 miles. Since the cumulative frequency cumulates to the end of the interval, one plots the cumulative frequency polygon by taking points at the ends of the intervals.

Cumulative Percentage Polygon. Another very useful technique is to plot the cumulative *percentages* rather than the cumulative *frequencies*. The cumulative percentages for our distance data are shown in column four of Table 2.3.1. Plotting at the ends of the intervals, the cumulative percentage polygon is shown in Figure 2.3.1.
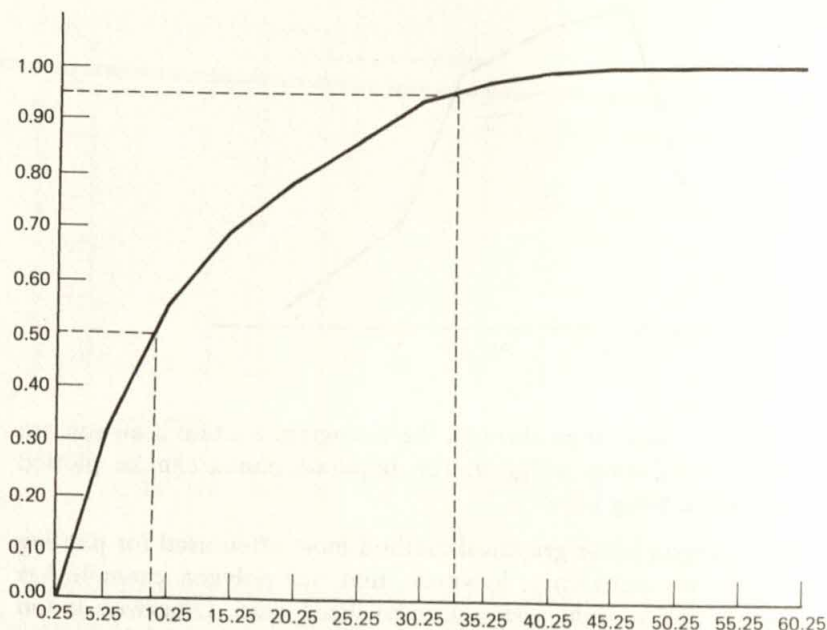
FIGURE 2.3.1 Cumulative percentage polygon of commuting distances (miles) of upper-class community college students—Mecca Community College.

By using this graph we can read off any percentage point we would like. For example, we can find the 50th percentage point. Tracing across from .50 to the polygon, and then following a vertical straight line from that point to the horizontal axis, we find that 50 percent of the students in our sample of $n = 180$ travel less than 9 miles and the other 50 percent of the students travel more than 9 miles.

We can also determine a completion to the statement: 95 percent of the students travel no more than ? miles. Reading this off the curve, we find it to be 33 miles.

EXERCISES

2.3.5 In an article [Welton, D. G., Inside dermatology, U.S.A.—from a national survey of private office practice, *South. Med. J. 53* (2), 210–223 (1960), Table 2], the following data on patients were shown.

Age and Sex of Patients Seen During Average 4-Week (Interrupted) Period

Age	Male	Female	Total
Under 10	16	19	35
10–19	41	59[a]	100
20–39	62	100	162
40–49	34	45	79
50–69	55	66	121
70+	16	16	32
Total	224	305	529

[a] Number modified for our purpose

a. Draw a histogram for each sex showing the age distribution of patients.

b. Using a frequency polygon, show the age distribution of the total patient load.

c. Using a cumulative percentage polygon, determine the following: (a) under what age 90 percent of all patients are, (b) below what age 90 percent of all women patients are, and (c) what age category is the most prevalent (justify your answer).

2.3.6 In the example on carcinoma of the cervix (page 44), the author stated ". . . it will be noted that almost a majority of the patients fall into the age group 40–55, commonly known as the cancer age."

a. Discuss the correctness of this statement, using the numerical data and the graphs that have been constructed in the text.

b. Is the author's age grouping meaningful? Why?

c. Is the grouping of ages convenient for rapid comparisons by age? If not, what would you recommend?

2.3.7 In a study [Greene, G. R., and Sartivell, P. E., Oral contraceptive use in patients with thromboembolism following surgery, trauma, or infection, *Am. J. Public Health* 62 (5), 680–685 (1972), Table 2], the age at hospitalization distribution of the cases and controls responding to a questionnaire were shown as follows: *

Age at Hospitalization	Cases		Controls	
	Responding	Nonresponding	Responding	Nonresponding
15–19	5	6	6	3
20–24	17	6	17	13
25–29	8	4	13	14
30–34	13	6	7	14
35–39	12	7	16	11
40–44	5	4	1	3

a. Using graphical methods, compare the age distribution of the case respondents to the control respondents.
b. Compare the age distribution of cases and controls among the nonrespondents.
c. Is there a difference between cases and controls relative to their disposition to respond to the questionnaire? Justify your response graphically.

2.3.8 In the study referred to in Exercise 2.3.7 above, another table showed the "education-grade completed" for the people responding: **

Education Grade Completed	Cases	Controls
8 or less	5	1
9–11	4	4
12	14	21
13–15	22	20
16	6	8
17 or more	9	6

a. Show in graphical form the education relationship tabulated here.
b. Is there anything misleading in these tables? State why your graph or graphs clarify the situation.

2.3.9   In the *Chicago Tribune* (Monday, May 8, 1972, Section 1, p. 2), the following table on West Side Poverty was printed.

**Poverty On The West Side**
(U.S. Census Bureau Study of Low-Income Areas[a])

	Families with Male Head			Families with Female Head		
	Negro	Latin	White	Negro	Latin	White
Total persons	65,146	27,557	56,169	84,087	24,059	55,213
Those with incomes below poverty level	13,741	4,277	9,429	28,400	6,101	11,036
Percent below poverty level	21%	16%	17%	34%	25%	20%
Poverty level unemployed	6,971	1,340	2,865	6,543	1,112	1,355

[a] The area surveyed and reflected in the statistics included most of the West Side as well as some economically depressed areas of the North Side in which large numbers of Negroes, Latins, and persons of Appalachian background live. The survey was taken following the 1970 census.

   a. Using the data in this table, present a picture of the poverty on the West Side of Chicago by graphical techniques. Use more than one kind of graph.
   b. What can you say about the growing poverty problem on the West Side?

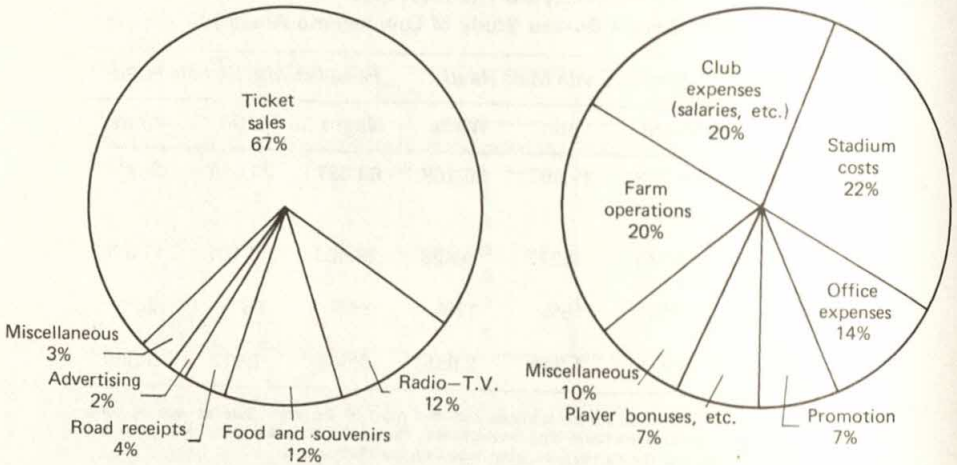2.3.10   The following graph appeared in a newspaper article.
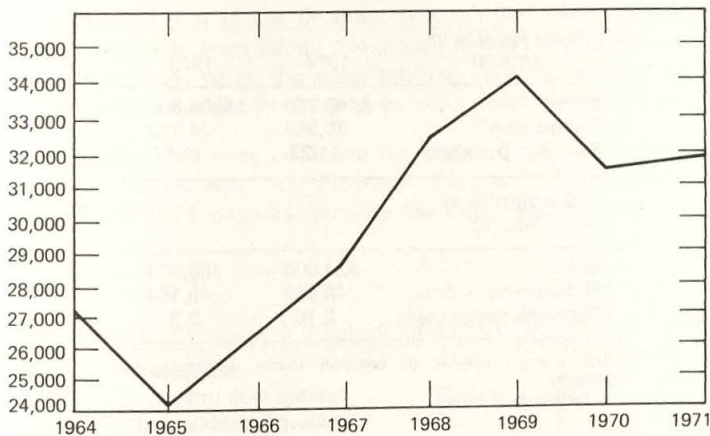
   a. What conclusions can you draw from this graph?
   b. Criticize the graph. State its limitations.
   c. What recommendations would you make to improve the graph?

2.3.11 The following graph on a "Baseball Team Budget" appeared in the *New York Times* on April 9, 1972.*

Where it comes from. . . . .   Where it goes . . . .

Ticket sales 67%

Miscellaneous 3%
Advertising 2%
Road receipts 4%
Food and souvenirs 12%
Radio—T.V. 12%

Club expenses (salaries, etc.) 20%
Stadium costs 22%
Farm operations 20%
Office expenses 14%
Miscellaneous 10%
Player bonuses, etc. 7%
Promotion 7%

Anatomy of a baseball-team budget.

a. Discuss the positive and negative aspects of this graphical presentation.
b. Show how you would have drawn the two graphs, and justify any changes you made.

2.3.12 The following table shows the progress of a major U.S. corporation.

Three Months to June 30	1974	1973
Sales[a]	$447,700	$384,800
Net Income[b]	25,359	24,173
Earnings per share	1.22	1.17

Six Months to June 30		
Sales[a]	834,000	755,000
Net Income[b]	44,869	48,154
Earnings per share	2.16	2.32

[a] The average number of common shares outstanding is 20,785,858
[b] In thousands of dollars

a. Draw a graph showing the rise in earnings per share for this corporation in these time intervals.
b. If you were a stockholder, would you be concerned about the progress of this company? Why?

2.3.13* The following excerpt from an article, "Continued Job Declines Threaten City Economy," appeared in the Sunday edition of the *New York Times*, July 21, 1974. Using the information in the article, show graphically any facts you would like brought out.

New York City is losing at an accelerating rate the jobs that sustain its economy and its government, and the declines are causing growing concern among city officials, private economists, business men and labor leaders.

There were, on the average, 24,000 fewer jobs in the first four months this year than last year.

Leading the declines are jobs in manufacturing. These are just the jobs that are most needed here to provide entry opportunities for the city's increasing population of poor and unskilled Puerto Ricans, blacks and other minorities.

Though other American cities also are losing jobs, and some of them at a faster rate than New York, the problem here is more serious because New York is bigger and therefore is losing more—251,000 jobs in the last four years.

The decline, which manifests itself in empty lofts and factory buildings, a high rate of unemployment (7 percent locally compared with 5.2 percent nationally) and a huge burden of welfare dependency, is confirmed by an array of statistical evidence that has been pointing downward since 1969.

After a decade of employment growth in the nineteen-sixties, the city lost 53,000 jobs in 1970, 135,000 in 1971, 49,000 in 1972 and 14,000 in 1973. The losses wiped out all the gains achieved in the previous decade.

By far the largest part of the decline—169,000 jobs—was in manufacturing employment, which had been falling continually through the fifties and sixties as well. However, in those years, the manufacturing losses were made up and even exceeded by growth in office work, services and government employment.

2.3.14 The following data were published by Rinfret Boston Associates, Inc., 1974, *Prices and Production, An Economic Analysis of Softwood Lumber and Plywood, 1970–73*, page 54, showing the distribution of softwood in 1970.**

	Lumber (percent)	Plywood (percent)
Residential construction market	50	59
Nonresidential market	9	11
Industrial market	20	16
All other markets	21	14

Organize these data into pie charts.

* ©1974 by The New York Times Company. Reprinted by permission.

**Reprinted with permission of the North American Lumber Association, Inc.

2.3.15 In the pamphlet "Pneumoconiosis in Coal Miners," Public Health Service Publication No. 2000, Figure 20 shows a three-dimensional plot of the relationship between age and years worked underground on roentgenographic findings of definite pneumoconiosis among working and nonworking miners. Write down your assessment of this graphical technique and draw any inferences from the figure you would like to make.

Working miners   Nonworking miners
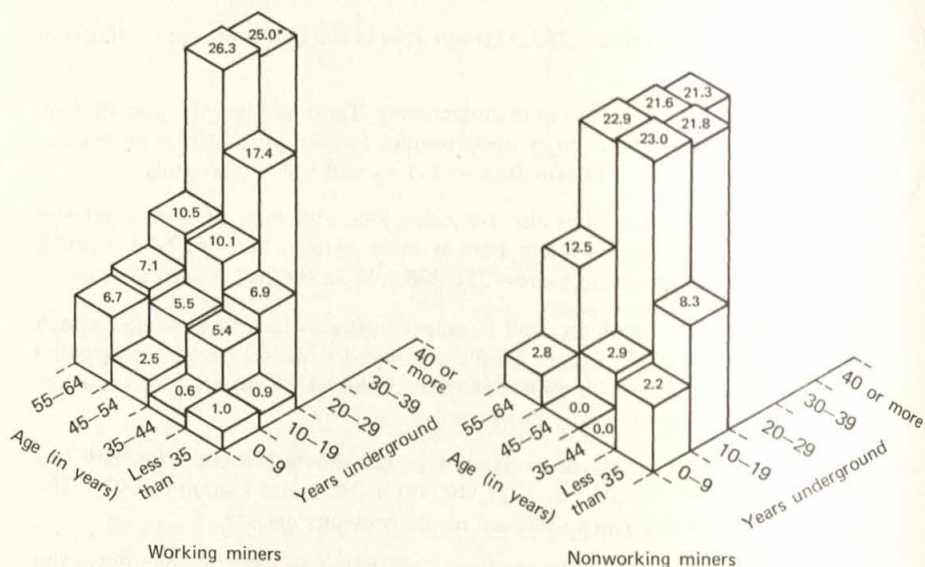
Roentgenographic findings of definite pneumoconiosis by age and years underground among working and nonworking miners.

* 25.0 = the percentage of 55–64 year old miners who had worked more than 40 years underground and had definite pneumoconiosis.

2.3.16 Using the following data and the previous exercise as an example of a three-dimensional plot, make a similar chart relating smoking behavior and years worked underground with history of persistent coughing.

Years Underground	Degree of Cough	Nonsmokers and Previous Smokers	Present Smokers
	None	182	233
0–9	<Persistent[a]	22	57
	Persistent	5	35
	None	177	297
10–19	<Persistent	20	74
	Persistent	3	28
	None	209	286
20–29	<Persistent	41	113
	Persistent	16	74
	None	187	141
30–39	<Persistent	34	89
	Persistent	16	45
	None	57	30
40+	<Persistent	14	19
	Persistent	4	18

[a] <Persistent = less than a persistent cough

2.3.17 The distribution of the U.S. land mass in January, 1970 is shown in the following table taken from the U.S. Department of Agriculture, *Outlook for Timber in the United States*, October 5, 1973, pp. 225–226. Draw a graph or chart of these data, explaining why you chose the particular graph you used.

U.S. Land Mass[a]
(January 1, 1974)

	Million Acres	Percent of Total
Crop land	427.0	18.8
Commercial forest	499.7	22.0
Noncommercial forest	253.9	11.2
Unproductive	233.9	10.3
Productive reserved	17.2	0.8
Deferred	2.7	0.1
Other lands	1089.5	48.0
Total	2270.1	100.0

[a] Modified for class use.

2.3.18   The following graph* on the cost of living in ten countries appeared in the *Saturday Review/World* on July 27, 1974. Interpret in words what you see in the graph. Criticize the graph constructively. **

Year	Money supply ($ Billions)	Index of cost of living (1963 = 100)
1958	$217	90.0
1959	225	91.0
1960	232	92.8
1961	249	94.7
1962	266	97.2
1963	289	100.0
1964	307	102.5
1965	332	105.5
1966	348	109.3
1967	378	112.4
1968	415	117.0
1969	441	112.6
1970	485	129.6
1971	550	136.4
1972	625	142.4

Ten Countries [a]

Belgium	Netherlands
Canada	Switzerland
France	United Kingdom
Germany	United States
Italy	
Japan	

[a]Chart shows total money—supply increase of ten major countries with corresponding increase in cost of living index, 1958—1973.

* *Source:* International Monetary Fund Financial Statistics, various issues. Compiled by Sidney E. Rolfe from issues of the International Monetary Fund *Financial Statistics.*
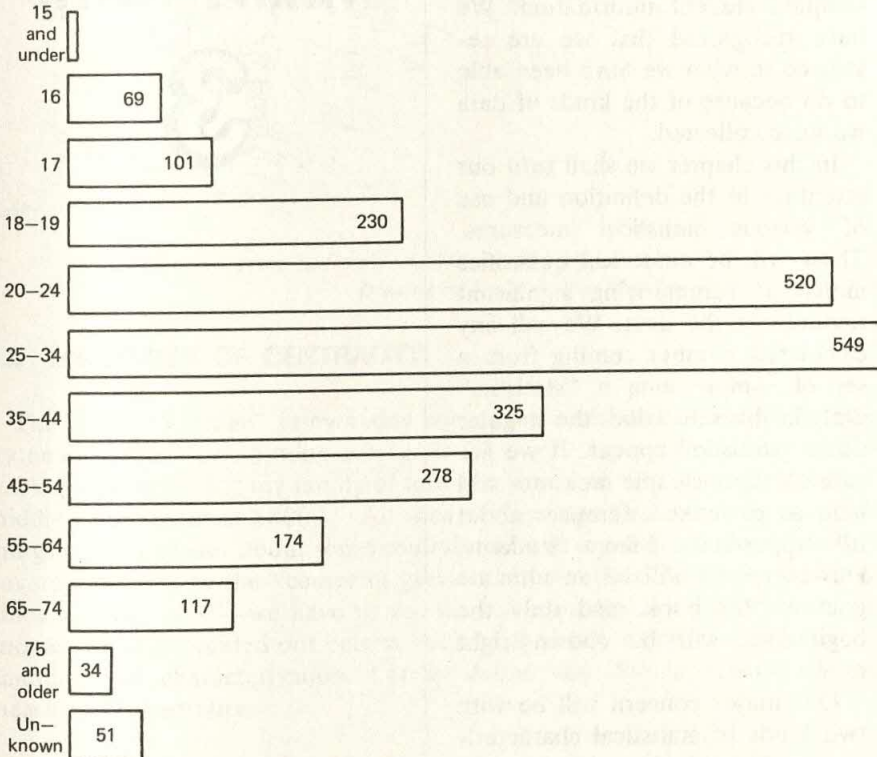**Reprinted by permission from Saturday Review/World, © 1974.

2.3.19 The following graph appeared in the Raleigh, North Carolina *News and Observer*, Sunday, May 6, 1973:

The chart shows the number of drivers in different age groups involved in North Carolina fatal crashes in 1972.

a. Using this graph, what implications and conclusions would you draw about driver's age and fatalities?

b. How would you correct the graph? What conclusions would then be drawn?

Drivers by age

Age	Number
15 and under	
16	69
17	101
18–19	230
20–24	520
25–34	549
35–44	325
45–54	278
55–64	174
65–74	117
75 and older	34
Un-known	51

3.1 INTRODUCTION

We began our study with a sample of $n = 180$ observations on upper-class students at Mecca Community College. Our objectives are to characterize or describe these students as best we can, and on that basis to make some statements about *all* the upper-class students at Mecca Community College. In Chapters 1 and 2 we have concentrated on tabular and graphical representations of our sampled student information. We have recognized that we are restricted in what we have been able to do because of the kinds of data we have collected.

In this chapter we shall turn our attention to the definition and use of various *statistical* measures. These will be numerical quantities aimed at summarizing significant features of the data. We call any calculated number coming from a set of sample data a "statistic." Only in this sense does the singular noun "statistic" appear. If we are careful, these sample measures will help us to make inferences about *all* upper-class Mecca students. However, this will be an ultimate goal of this book, and only the beginnings will be shown right now.

Our major concern will be with two kinds of statistical characterization of sample data: (a) measures of centrality, and (b) measures of variability.
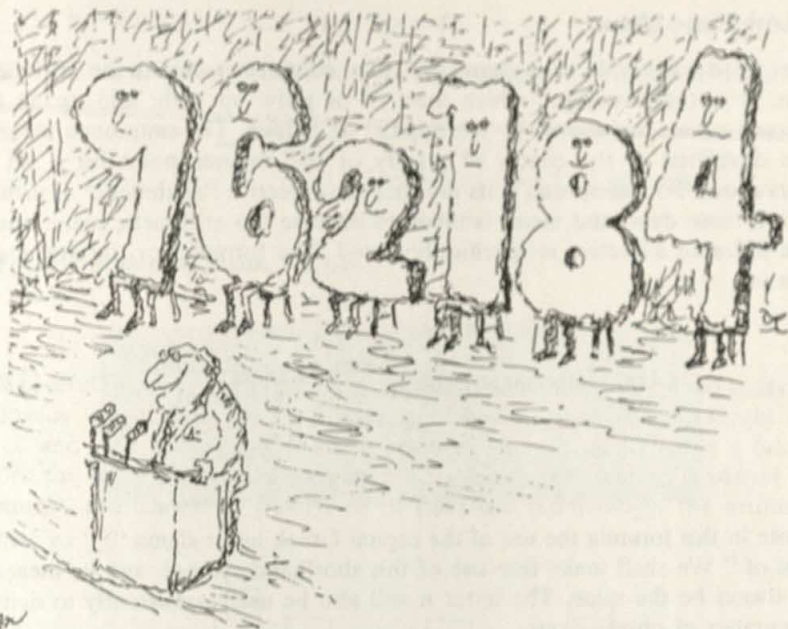
3

Summarizing Data Numerically

3

"*Tonight, we're going to let the statistics speak for themselves.*"
Drawing by Koren; © 1974 The New Yorker Magazine, Inc.

3.2 MEASURES OF CENTRALITY

The use of "average" in everyday conversation abounds. Everybody makes statements like "Joe's batting average is .303," or "I spent an average of $34 a week for groceries for my family of four last year," or "the average income for middle-class whites is $8000." All uses of "average" refer to a concept that things tend to cluster about some central value. In some fields like science, the "average" refers to the "center of gravity." In statistics, we recognize several kinds of "averages"—we have to do this since we deal with all kinds of data and, as we have pointed out before, each kind of data requires special ways of handling and characterization. Let us define and discuss several of these measures of centrality.

A. Arithmetic Mean

The most commonly used characteristic of continuous data is the arithmetic mean. It is the "average" which most of us grow up with: add up all the observed values and divide by the number of entries. The arithmetic mean is often described as the center of gravity or the balance point for a set of observations. So widespread is its use that the adjective "arithmetic" is usually omitted these days, and *mean* is taken to indicate the arithmetic mean unless some different adjective is specifically stated. The formula for the arithmetic mean is:

$$\text{(Arithmetic) mean equals } \frac{\text{Sum of observations}}{\text{Number of observations}} \tag{3.2.1}$$

$$\bar{y} = \frac{\sum y}{n}$$

Note in this formula the use of the capital Greek letter sigma "$\sum$" to denote "sum of." We shall make free use of this shorthand symbol, and its meaning will always be the same. The letter n will also be used consistently to denote the number of observations.

A more precise way to indicate the arithmetical operations used in Formula (3.2.1) is to write it as follows:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.2.1'}$$

The subscript i on the y indicates that y: y_1 is the first y, y_2 is the second y, and so on. The complete $\sum$ notation then indicates that y is to take successively each of the values from the first through the nth value and these n values are then to be added. This complicated notation to symbolize the simple process of addition may seem unnecessary at the moment, but overall it is a great time and space saver. Let's illustrate this with an example.

Example 3.2.1

Suppose we have the following seven sample observations of weights (in pounds):

$y_1 = 1$ pound $\qquad$ $y_2 = 4$ pounds $\qquad$ $y_3 = 6$ pounds $\qquad$ $y_4 = 1$ pound

$y_5 = 6$ pounds $\qquad$ $y_6 = 2$ pounds $\qquad$ $y_7 = 1$ pound

The sum of the seven sample observations is denoted in complete detail by:

$$\text{Total} = \sum y = \sum_{i=1}^{7} y_i = y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7$$
$$= 1+4+6+1+6+2+1 \text{ (pounds)}$$
$$= 21 \text{ (pounds)}$$

For the *arithmetic mean,* we have

$$\bar{y} = \frac{\sum y}{n} = \frac{21}{7} = 3.0 \text{ (pounds)}$$

A useful way of understanding the arithmetic mean is to consider the following weight scale with the number of observations at each weight indicated and look for a pivot point to balance the scale. If we place a fulcrum under the scale so that the weights are balanced, the fulcrum is placed at 3 (pounds), the center of gravity, or in statistical terminology, the arithmetic mean.

y (pounds)
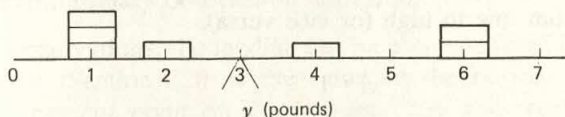
There is one notable peculiarity of the arithmetic mean: it is strongly influenced by a few odd-ball observations. Suppose that the sample of Example 3.2.1 has one more observation added and the additional observation turns out to be $y_8 = 35$ pounds. Then the sum of all y values is $21+35 = 56$, and $\bar{y} = (56/8) = 7.0$. Thus the mean has gone from 3 to 7 pounds just because of one extraordinary observation. You can imagine the change in the mean per capita income of your town if the world's richest man should move in.

The (arithmetic) mean has the great advantages of being easily computed and readily understood. You cannot use it, of course, to average hair colors or political party preferences. But for numerical observations, it is the average that gets first attention. For continuous data, it gives a definite realizable value; a mean weight $\bar{y} = 6.849$ pounds can exist. For discrete data, it often has to be understood as hypothetical: if weekly numbers of telephone calls over a year give $\bar{y} = (647/52) = 12.44$, we cannot actually have forty-four one-hundredths of a phone call—except "on the average." If measurements of opinion on a one-to-five scale are made and lead to $\bar{y} = 3.78$, we have to say that the mean is somewhat below scale 4.

(Arithmetic) Mean Summary. The primary characteristics of the (arithmetic) mean are as follows:

1. It is used for all continuous measurements, and, with restricted meaning, for ordinal data.

2. Each observation (measurement) in the sample is included in the calculation.

3. Extreme (very large or very small) observations have a heavy effect on its numerical value.

B. Median

Another important measure of centrality is called "the median." If we arrange the measurements from lowest to highest and select the "middle one," we have what is known as the *median*. If we have an odd number of observations, the median is exact. If we have an even number, we must average the "two middle" values.

Example 3.2.2

Using the sample of seven weights in Example 3.2.1, we must first arrange the weights from low to high (or vice versa).

$$y_1 = 1$$
$$y_4 = 1$$
$$y_7 = 1$$
$$y_6 = 2 = \text{middle value}$$
$$y_2 = 4$$
$$y_3 = 6$$
$$y_5 = 6$$

The value exactly in the middle is $y_6 = 2$ (pounds). Therefore the median $= 2$ (pounds).

In actual practice we seldom need to carry along the y designations (y_1, y_2, y_3, etc.) in any calculation or summary. If any designation is called for, it is more likely to be something for indicating the place of the observation in a ranking by size. For that purpose, subscripts enclosed in parentheses are used. For instance, the data of Examples 3.2.1 and 3.2.2 are as follows:

$$y_1 = 1 = y_{(1)}$$
$$y_4 = 1 = y_{(2)}$$
$$y_7 = 1 = y_{(3)}$$
$$y_6 = 2 = y_{(4)} = \text{middle value}$$

$$y_2 = 4 = y_{(5)}$$
$$y_3 = 6 = y_{(6)}$$
$$y_5 = 6 = y_{(7)}$$

Note what happens to the median here if the outlying y_8 is added to the sample, as we discussed with reference to the mean. We would have:

$$y_1 = \ 1 = y_{(1)}$$
$$y_4 = \ 1 = y_{(2)}$$
$$y_7 = \ 1 = y_{(3)}$$
$$\left. \begin{matrix} y_6 = \ 2 = y_{(4)} \\ y_2 = \ 4 = y_{(5)} \end{matrix} \right\} = \text{two middle values}$$
$$y_3 = \ 6 = y_{(6)}$$
$$y_5 = \ 6 = y_{(7)}$$
$$y_8 = 35 = y_{(8)}$$

The median is now $(2+4)/2 = 6/2 = 3.0$, not a very large change from what it was before the extraordinary observation appeared.

For any set of observations, the median can be obtained from the cumulative percentage polygon (Section 2.3) as the value on the horizontal axis determined by the 50 percent point on the polygon. You will recall that we did this in Figure 2.3.1.

Median Summary. The primary characteristics of the median are as follows:

1. The median can be used with the ordinal type of discrete data and with both types of continuous data (interval and ratio scale).

2. Once the data have been ordered from either low to high or high to low, the median ignores all the observations except the one (or two) in the middle of the ordered array.

3. Extreme observations have very little effect on the median.

4. The median is the best average to use for any characteristic which by its nature in an entire population has many more extreme values on one side of the arithmetic mean than on the other. Such a distribution is said to be badly (or severely) *skewed*. Family income, per capita wealth, and annual wages are common examples of skewed distributions. A few extremely wealthy people in a community can cause the arithmetic mean income to be far above the typical income unless we state *average* in another way. In such a case, the median income is usually considered the appropriate measure of central tendency.

C. Midrange

The midrange is defined as the arithmetic mean of the lowest and highest observations in the sample:

$$\text{Midrange} = \frac{\text{lowest value} + \text{highest value}}{2} \qquad (3.2.2)$$

In the first example on weights, the smallest weight is 1 pound and the largest weight is 6 pounds. Thus

$$\text{Midrange} = \frac{1+6}{2} = 3.5 \text{ (pounds)}$$

Midrange Summary. The basic characteristics of the midrange are as follows:

1. The midrange is used only with interval or ratio scale data.

2. It is an easily determined value and is a very efficient estimate when the sample size is small.

3. Extreme observations have a large effect on the midrange.

D. Mode

The most frequently occurring value among a group of sample observations is called the *mode*. It appears as a peak in the graph of the frequency distribution of observations.

In viewing distributions, it is wise to ascertain that only one peak occurs. If two peaks occur, any single measure of central tendency may give a misleading impression by implying that it describes the more common single-mode type of data. We call such a two-peaked distribution "bimodal."

Example 3.2.3

Using the seven sample weights of Examples 3.2.1 and 3.2.2, we see that the 1-pound weight occurs 3 times, the 6-pound weight occurs twice, and the 2-pound and 4-pound weights occur only once. Therefore the modal value is one pound; thus

$$\text{mode} = 1 \text{ pound}$$

Mode Summary. The basic characteristics of the mode are as follows:

1. The mode can be used for nominal, ordinal, interval, and ratio types of data.

2. In frequency distributions like the one determined for "commuting distance" in the discussion on histograms, the modal value is the midpoint of the interval that contains the greatest number of observations.

3. Extreme observations have little effect on the mode.

E. Other Measures of Centrality

There are two other special kinds of average that are sometimes discussed: the geometric mean and the harmonic mean. These have use in certain very specialized situations that arise rather rarely in ordinary practice. Since the remainder of our text does not utilize these measures, they are omitted from the book.

A Review Illustration.   *Example 3.2.4.*

The following sample of $n = 9$ observations was taken from the set of data recorded in the table under Summary Exercise 3.5.1.

$$y_1 = 2400 \qquad\qquad y_6 = 3650$$
$$y_2 = 2750 \qquad\qquad y_7 = 2180$$
$$y_3 = 2180 \qquad\qquad y_8 = 2000$$
$$y_4 = 2320 \qquad\qquad y_9 = 2190$$
$$y_5 = 1930$$

Using these data, calculate the following: (a) mean, (b) median, (c) mode, and (d) midrange.

(a) Mean [using Formula (3.2.1)]:

$$\bar{y} = \frac{\sum y}{n} = \frac{\text{total}}{n},$$

$$\bar{y} = \frac{21,600}{9} = \$2400.$$

The center of gravity (the equilibrium point) is $2400.

(b) Median:

First arrange the observations in order from lowest to highest. The value of the observation that is in the center of this ordered set of observations is the median:

$1930	2320
2000	2400
2180	2750
2180	3650
2190 = median value	

(c) Mode:

The observation which occurs most frequently is the modal value. By inspecting the above set of data, we find the mode = $2180. (In small samples of wide-ranging data, it is highly unlikely that there will be two observations alike. In such cases, we have to report that the sample "has no mode.")

(d) Midrange:

The midrange as defined by Formula (3.2.2) is calculated as follows:

$$\text{Midrange} = \frac{\text{smallest observation} + \text{largest observation}}{2}$$

$$= \frac{1930 + 3650}{2} = \frac{5580}{2} = 2790,$$

Midrange = $2790

Discussion of Review Illustration. Analysis of the data has resulted in the following:

1. Mean = $2400
2. Median = $2190
3. Mode = $2180
4. Midrange = $2790

All these statistics are correct, but if one is determined to make inferences about the *average* income level of these counties, a choice among them needs to be made.

The *mean* is the only statistic that uses *all* the data. Every data point has equal weight in calculating $\bar{y}$. In general, this is the best of all the averages, *but* there should be some doubt in this case. Why? There is one very large income-data point, namely 3650, and this observation is making the mean large.

The *median* is concerned only with the observation that is in the middle of the ordered set of observations. Thus it would make no difference how large the largest one was, or how small the smallest; the median would not change. For example, suppose that the above data were:

	Original data	New data (1)	New data (2)
	$1930	$1930	$ 140
	2000	2000	2000
	2180	2180	2180
	2180	2180	2180
Median	2190	2190	2190
	2320	2320	2320
	2400	2400	2400
	2750	2750	2650
	3650	6350	2750
Arithmetic mean	2400	2700	2090

In all these cases the median remains unchanged, while the arithmetic mean changes a great deal.

The *mode* is an interesting statistic but is concerned with only the most frequent observation. This is not a useful statistic in small samples. The concept of modality *is* important when we discuss large distributions.

The *midrange* uses only the two extreme observations and is thus subject to extremeness in small samples, particularly here.

Thus the choice is between the mean and the median. In income statistics the median is usually used since it is deemed to be more representative of the underlying income distribution. However, if you are truly interested in a center of gravity type of statistic, the mean would be used.

EXERCISES

3.2.1 Twenty bottles of 14-ounce "Misty" mouthwash were taken from a grocery store shelf, and the amount of liquid in each bottle was measured in cubic centimeters (cc) and recorded as follows:

Sample	Volume (cc)	Sample	Volume (cc)
1	423	11	422
2	426	12	422
3	421	13	426
4	428	14	422
5	420	15	430
6	418	16	425
7	423	17	426
8	426	18	420
9	427	19	418
10	428	20	421

a. Determine the (arithmetic) mean volume in these 20 bottles.
b. Determine the median volume of the 20 bottles.
c. What is the modal volume of the 20 bottles?
d. Calculate the volume that represents the midrange of the 20 bottles.
e. As a result of the above calculations, what can you say about the symmetry of the volumes of the liquid in the 20 bottles?

3.2.2 A chemist was working with a new chemical reaction, and he was concerned about the length of time it took for the reaction to be complete. He decided to take 12 different samples of raw material and time the length of reaction for each sample. He recorded (columnwise) the following times:

17	12	11	10	10	9
14	11	10	10	9	9

a. What was the average time to completion? Discuss your answer and justify whatever measure of central tendency you would use.
b. If the data above were recorded in the same order that the samples were used, would you be concerned about the data? Why or why not?

3.2.3 The amount of total synthetic detergent (percent) in a soap product can be measured through a chemical analysis known as "cationic" SO_3 titrations. The results obtained from 24 samples of a particular detergent from a store shelf were as follows:

36.8	36.8	36.3	36.1	35.0
36.6	35.3	36.4	36.4	35.4
36.8	36.2	37.5	36.2	35.4
36.8	36.9	36.3	36.1	35.3
36.1	36.3	36.3	35.6	

a. Determine the arithmetic mean of the data.
b. Determine the median.
c. Determine the mode.
d. If the manufacturer claimed that his product contained less than 38 percent, would you be concerned about his claim?
e. If the manufacturer claimed that his product contained less than 37 percent, would you be concerned about his claim?
f. Do you think it is possible for a manufacturer to guarantee that every box of detergent he produces has less than 37 percent TSD? (Assume, for discussion purposes only, that the product must have 35 percent TSD in order to be a "good" cleaner.)

3.2.4 Each sales district of a major sweater company reports its sales figures at the end of each quarter. The company has established a sales quota (target) for each district for each of the quarters of the year. The following data represent the position of each sales district relative to its quota as of July 1, 1971.

District Number	Percentage of Quota	District Number	Percentage of Quota
1	82	11	105
2	95	12	110
3	115	13	126
4	104	14	110
5	96	15	109
6	118	16	88
7	110	17	96
8	98	18	102
9	84	19	98
10	104	20	108

a. Summarize the sales status of the company for the sales vice-president.
b. What proportion of the sales districts had met their established quota?
c. What proportion of the districts were doing better than 110 percent of their quota?

d. If the company's sales quota for the half year was 100,000 sweaters, assigning 5000 to each district, did the company meet its goal? How many sweaters did the company sell? Suppose that each of the eight districts that sold 108 percent of quota or better had a quota of 2000 sweaters, while each of the other 12 districts had a quota of 7000. Then what was the total of the company's sales? What warning do these results give you about totaling or averaging percentages?

3.2.5 A new liquid fabric finish is sold in a plastic bottle with a cap measured to hold about 1.4 ounces of liquid. The instructions on the bottle state that one capful is needed for a normal load of wash. Suspecting that some housewives spill in extra product, while others do not use enough, the manufacturer obtained a sample of usage measurements for ten housewives.

Ounces of Product Used	
1.1	1.4
0.8	1.2
1.2	1.6
1.0	1.3
0.9	1.5

a. As a result of these 10 measurements, what conclusions would you draw about the usage of the product?
b. Would you conclude that housewives do not follow directions? Why or why not?
c. How many usage measurements would you feel necessary before you would be willing to state that housewives, on the average, use more product than the instructions state?

3.2.6 The American Kennel Club (A.K.C.) registers over 1,000,000 pure-bred dogs each year. There are 116 different breeds recognized by the A.K.C. Each breed is ranked according to the number of dogs that are registered. In 1971 the toy breeds had the following ranks:

1	41	91
6	42	100
8	60	109
16	67	53
17	84	45
26	87	

a. What is the mean rank of the toy breeds in 1971?
b. What is the mode? What is the median?
c. Which statistic would you use to represent the average ranking of toy dogs in 1971? Why?

3.2.7 During the week ending January 30, 1971, the fifteen most active stocks on the New York Stock Exchange reported the following information:

Most Active Stocks

Company	Volume	Final Price	Net Change
East. Air Lines	1,014,600	$20\frac{3}{8}$	$+2\frac{7}{8}$
Texaco	987,400	34	$+\frac{5}{8}$
Sperry Rand	958,300	$28\frac{5}{8}$	$+1\frac{1}{2}$
Fed. Nat. Mtg.	886,200	61	-2
Trans. W. Air.	836,600	18	$+1\frac{1}{2}$
Am. Airlines	813,500	$28\frac{1}{4}$	$+1$
Nat. Cash Reg.	751,200	$39\frac{1}{2}$	$+\frac{1}{2}$
Pan Am.	718,500	$16\frac{1}{8}$	$+1\frac{7}{8}$
Tex. Gulf Sul.	616,500	21	$+2\frac{7}{8}$
A.T. & T.	559,700	$53\frac{1}{2}$	$+1\frac{1}{4}$
I.T. & T. pf N	554,500	69	$+\frac{3}{4}$
Occidental Pet.	552,500	$18\frac{3}{4}$	$+1\frac{1}{4}$
Gulf Oil	529,200	$29\frac{7}{8}$	$-\frac{1}{8}$
Nwst. Airlines	518,700	$27\frac{3}{8}$	$+2\frac{1}{8}$
Mad. Sq. Gar.	505,300	$5\frac{1}{8}$	$+1$

a. Calculate the arithmetic mean and the median for volume, last price, and net change.
b. Do you believe that these 15 stocks would represent the general condition of the market during this week? Why or why not?
c. Obtain a copy of the Sunday *New York Times* and determine last week's most active stocks. Do the same calculations that were done in (a). What conclusions can you make?

3.2.8 The *New York Times* Weekly combined averages for the New York Stock Exchange are published each Sunday in the *Times*. One is shown on page 73.
a. Using the data shown on the graph, answer the following questions:
   i. What percentage of the days had a closing price closer to the low for the day than to the high for the day?
   ii. In the six lowest sales-volume days, does the closing price tend to be closer to the low or to the high for the particular day?
   iii. Using this graph, estimate the average daily sales volume. What statistic would you use?
   iv. What was the average daily closing price on the New York Exchange in the months shown?
b. If you had all the measures of central tendency calculated, would these be sufficient to describe these data? Why or why not?

3.2.9 A food company produced a new product called "Pizza Sticks," which were really baked shells filled with pizza. A sample of nine baked shells (empty) were taken from each of two different production lines and each shell was weighed. The weights were recorded as follows:

Weights in Ounces

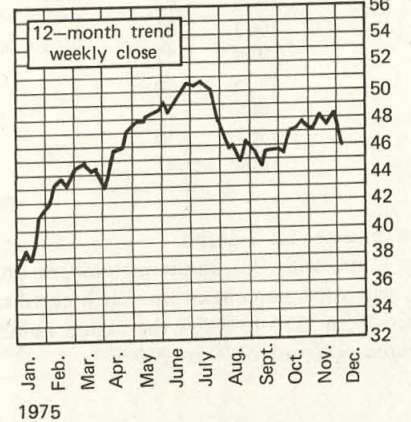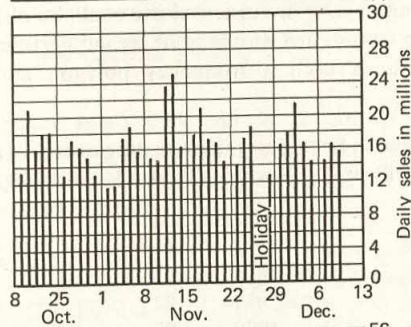| Line Number 1		Line Number 2	
1.14	1.22	1.27	1.22
1.28	1.17	1.33	1.19
1.23	1.16	1.20	1.30
1.20	1.30	1.22	1.22
1.22		1.25	

a. Calculate the average weight produced on line 1 and on line 2 using the following statistics: (i) arithmetic mean (ii) median, and (iii) mode.

b. Based on the results obtained in (a), can you say that the two lines are producing pizza stick shells with the same weight? Why or why not?

c. What is the average weight of the pizza shells using the data from both lines? Do this calculation two different ways using the arithmetic mean.

3.2.10 The annual "gross" incomes of 11 families were recorded as follows:

19,000	10,500
16,000	10,500
14,500	10,500
13,500	10,500
13,500	8,500
10,700	

a. Use the following statistical measures to determine an average income for the 11 families: (i) arithmetic mean, (ii) median, (iii) mode, and (iv) midrange.

b. Discuss these four results and indicate why each might be useful to the reader.

Day's			Year to date
Sales	Tuesday	Year ago	1975	1974
15,680,000	16,040,000	15,700,000	4,476,074,958	3,298,680,302

New York
Stock Exchange
composite index

← High
← Closing
← Low

Daily sales in millions

Holiday

8   25   1   8   15   22   29   6   13
Oct.   Nov.   Dec.

12—month trend
weekly close

Jan. Feb. Mar. Apr. May June July Aug. Sept. Oct. Nov. Dec.

1975
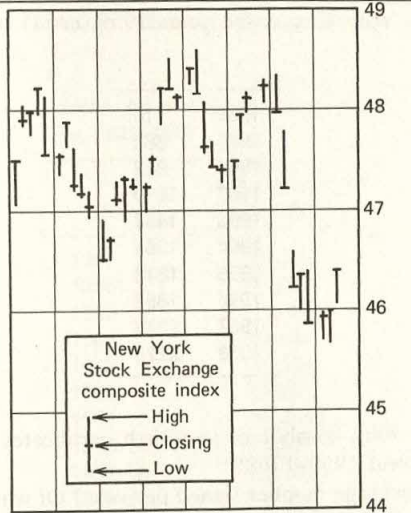
3.2.11 The number of live-birth certificates signed by Certified Nurse-Midwives (CNM) in New York City in the years 1959–1968* is shown below:

1959	10
1960	383
1961	912
1962	1610
1963	1493
1964	1861
1965	1819
1966	1852
1967	2278
1968	2225

a. What is the total number of live-birth certificates signed by CNM in the 10-year interval 1959–1968?
b. What is the average number signed per year? Of what use is this statistic?
c. What information of more practical use could be obtained from these data?
d. Tabulate the cumulative numbers of signed certificates in these ten years.
e. Plot the data as both a frequency polygon and cumulative frequency polygon.

3.2.12 In the same article indicated in the preceding exercise, a table reporting percent of live-birth certificates signed by CNMs, 1959–1968, was shown as follows:

1959	.01
1960	.23
1961	.54
1962	.98
1963	.89
1964	1.12
1965	1.15
1966	1.21
1967	1.56
1968	1.57

a. Plot these percentages on a graph.
b. Compare this graph with the frequency polygon in the preceding exercise. Explain why, when the percent of live-birth certificates signed by CNMs in 1968 increased from 1.56 to 1.57, the actual number of certificates signed by CNMs decreased from 2278 to 2225.

* Summary of Vital Statistics, The City of New York, C.N.Y. Department of Health, 1959–1968.

3.2.13 A table in a book [Cugliani, A., and Marano, P., *Heart, Cancer, Stroke, and Related Diseases* (1968), p. 48 (excerpted here for pedagogical use)] shows the following:

Deaths from Cancer, New York City, 1949–1967

Year	Total Number of Cancer Deaths
1967	17,788
1966	17,769
1965	17,402
1964	17,642
1963	17,254
1962	17,252
1961	17,384
1956–60	16,869
1952–55	16,553
1949–51	15,556

Source: Vital Statistics, 1949–67, New York City Department of Health.

a. What is the mean yearly total number of cancer deaths in these data?
b. Do you think this mean is a useful statistic? Why or why not?
c. Plot these data using a frequency polygon with the year on the *x*-axis.
d. What conclusions can you draw from these data?

3.2.14 * In an article [Robertson, Robert L., Economic effects of personal health services: Work loss in a public school teacher population, *Am. J. Public Health* 61 (1), 30–45, (1971), Table 1] are shown the mean days of work loss (rounded) in the study year 1966–1967 for a public-school teacher population by age, sex, and health plan. For our purposes the table has been modified.

	Males		Females	
Age	Blue Type	Group Practice	Blue Type	Group Practice
20–24	5.35	3.33	6.07	7.13
25–34	3.68	3.65	6.37	5.10
35–49	4.02	4.18	5.96	5.73
50–59	3.67	3.44	6.77	5.73
60–64	6.94	2.29	7.76	7.88

a. Using a histogram, plot the data for males with a "blue"-type health plan.
b. Using a histogram, plot the data for males with a group-practice health plan.
c. Using frequency polygons, plot on the same graph the data for females with a blue-type health plan and females with a group-practice health plan.
d. If you wish to compare two health plans, which graphical procedure would you use?
e. In order to determine the average days of work loss in this study population of teachers, would you add up all the data in the table and divide by 20? Why or why not?

3.2.15 In 1969, 18-year-old males were subject to the draft (service in the armed forces). The order in which they were called up was determined by a "lottery," *supposedly* completely at random. January dates were drawn out in the following sequence.

Lottery Number	Date	Lottery Number	Date
17	Jan. 15	101	Jan. 5
52	Jan. 25	118	Jan. 23
58	Jan. 19	121	Jan. 16
59	Jan. 24	140	Jan. 18
77	Jan. 28	159	Jan. 2
92	Jan. 26	164	Jan. 30
186	Jan. 21	280	Jan. 20
194	Jan. 9	305	Jan. 1
199	Jan. 8	306	Jan. 7
211	Jan. 31	318	Jan. 13
215	Jan. 4	325	Jan. 10
221	Jan. 12	329	Jan. 11
224	Jan. 6	337	Jan. 22
235	Jan. 17	349	Jan. 29
238	Jan. 14	355	Jan. 27
251	Jan. 3		

a. Calculate the average lottery number for males born in the month of January.
b. If the draw was a truly random one, you would expect that January's average lottery number would be around 183. On the basis of this figure, would you consider that the draw was not random? Why or why not?

3.2.16 The average lottery numbers for the 12 months in the 1969 draft lottery were as follows:

January	201	July	180
February	203	August	173
March	226	September	157
April	204	October	182
May	208	November	149
June	196	December	122

a. Using a bar chart, plot these figures.
b. What conclusions would you draw by inspecting these data?
c. Based on the above monthly averages, what is the grand average of all of the lottery numbers?
d. What is the true grand average of the lottery numbers? Explain the discrepancy between this result and the one you obtained in (c).

3.2.17 A manufacturing concern introduced a new product on the market and retail sales began on June 1, 1970. During the next 20 months the shipments made to all outlets were watched closely for indications of sales problems. The data obtained were as follows (figures are sales in thousands of cases):

1970	June	73		April	51
	July	82		May	53
	August	97		June	68
	September	101		July	81
	October	43		August	76
	November	59		September	73
	December	43		October	50
				November	53
1971	January	52		December	55
	February	44			
	March	44	1972	January	53

a. What is the mean monthly sales figure during this time period?
b. What is the median monthly sales figure?
c. Plot the data using time on the horizontal axis.
d. What statements would you be willing to make about this new product?
e. What do you expect to be sold in the month of February 1972? State your reasons.

3.2.18 Small transistorized radios were produced in a certain factory. This factory was a three-shift operation. Each shift a sample of 90 radios was checked and graded on a quality rating a, b, c, d, e, or f, with "a" being first quality and "e" and "f" being off-quality or rejectable. A certain day gave data as follows:

Shifts	Quality Rating					
	a	b	c	d	e	f
Day (8–4)	15	25	8	6	18	18
Evening (4–12)	17	29	8	15	7	14
Graveyard (12–8)	8	21	10	27	15	9

a. What statistic or statistics would you use to characterize the shifts?
b. What conclusions would you be willing to make about the relationship between quality and shift? Can you be certain about these conclusions?

3.2.19 In a community in North Carolina, the occurrence of hepatitis cases for 5 years was reported as follows:

Year	1961	1962	1963	1964	1965
Number of cases	7	10	8	44	11

Comment on each of the multiple-choice completions in the following statement. The average number of cases as determined by the median of the 5 years' experience is: (a) 16, (b) 8, (c) 44, (d) 10, and (e) 21.

3.2.20 Calculate: (a) arithmetic mean, (b) mode, and (c) median for the following data:

Diastolic Blood Pressure (mm Hg)
98
86
88
97
88
72
90
88

3.2.21 Calculate: (a) arithmetic mean, (b) mode, and (c) median for the following data on height, reported here to the nearest inch.

Height (in.)	Frequency
57	2
58	4
59	14
60	41
61	83
62	169
63	394
64	669
65	990
66	1223
67	1329
68	1230
69	1063
70	646
71	392
72	202
73	79
74	32
75	16
76	5
77	2

3.2.22 Isoniazid given orally for 20 weeks after the beginning of infection will prolong the life of rats suffering from leprosy. The following data show the survival time for a group of 10 rats after such treatment.

| Survival Time (weeks) |
|:---:|
| 51 |
| 53 |
| 67 |
| 70 |
| 70 |
| 72 |
| 73 |
| 79 |
| 84 |
| 88 |

a. Calculate the arithmetic mean, median, and mode of these data.

b. On the basis of these data, can you say that the use of isoniazid prolongs the life of rats suffering from leprosy? Why or why not?

3.2.23　In a large supermarket, the following sales of ultralarge packages of detergent were recorded:

Week	Sales	Week	Sales
1	22	13	23
2	18	14	16
3	18	15	24
4	22	16	19
5	24	17	18
6	25	18	21
7	28	19	22
8	28	20	17
9	20	21	22
10	19	22	16
11	23	23	17
12	21	24	20

a. Plot the weekly sales figures against week number using a frequency polygon.

b. Calculate the arithmetic mean of each set of 4 weeks' sales (i.e., weeks 1–4, 5–8, ..., 21–24).

c. Plot the means of part b as a frequency polygon. At what values of the horizontal scale are these means plotted? What assumption is made when you plot these means?

d. Discuss the plotting of weekly sales figures against 4-week average sales figures. When would you use both plots?

3.3 MEASURES OF VARIABILITY

In the previous section we were concerned with the various methods of characterizing the centrality of a set of data. However, many different sets of data could have the same measurement of centrality and still be very different. For example, while the following four distributions all have the same arithmetic mean (i.e., the same balance point), they are very different from one another. We say that the *dispersion, variability,* or *variation* of the data is different in the various sets. Thus if we are to make good predictions or estimates from observed data, we need to include in our analysis some measure of the dispersion of the data. We shall consider the four main measures of variation in current use: (a) *range,* (b) *variance,* (c) *standard deviation,* and (d) *coefficient of variation.*
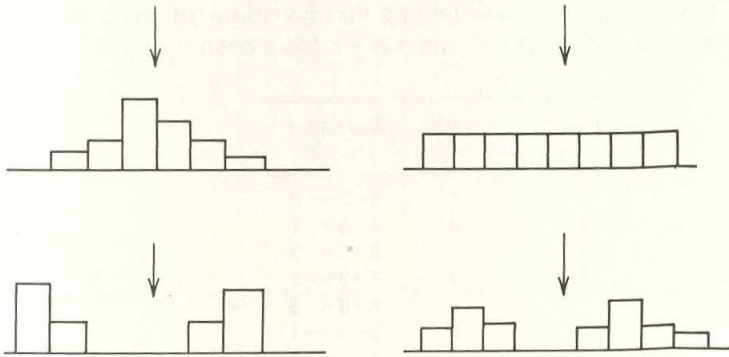
A. Range

The range is defined as the difference between the largest and the smallest observation.

$$\text{Range} = \text{largest } y - \text{smallest } y \tag{3.3.1}$$

Example 3.3.1

Using the seven weights used previously, where the largest weight was 6 pounds and the smallest weight was 1 pound, we calculate:

$$\text{Range} = 6 - 1 = 5 \text{ pounds}$$

The range is very easy to calculate, and it gives us some idea about the variability of the data. However, the range is at best a crude measure of variation, since it uses only two sample values. It is most useful with small samples, of size 10 or less.

B. Variance (s^2)

A better way to measure the variability of a set of data is to measure how each observation in the data set differs from the arithmetic mean $\bar{y}$ and then to obtain a statistic using these differences so as to reflect an "average" deviation from $\bar{y}$. Let us begin by determining the deviations of each weight from the mean weight in the sample of weights we have been using.

Observation Number	Weight y_i	Deviation $y_i - \bar{y}$
1	1	$1 - 3 = -2$
2	4	$4 - 3 = 1$
3	6	$6 - 3 = 3$
4	1	$1 - 3 = -2$
5	6	$6 - 3 = 3$
6	2	$2 - 3 = -1$
7	1	$1 - 3 = -2$
Total	21	0

$$\bar{y} = \frac{21}{7} = 3$$

Notice that the total of the column of deviations is *zero*. This is no mere coincidence. It will always be true for the sum of such deviations. After all, the mean $\bar{y}$ is the center of gravity of the set of data, and so the "overs" and "unders" have to balance out around that center.

Since the sum of the deviations from $\bar{y}$ is always zero, dividing this sum by n to get an average deviation would always yield zero also. Hence that would be of no use as a measure of the variability in the data. One mathematical device that makes sense to get around the difficulty is to square each deviation, add up

the results, take some kind of average, and extract the square root to get back to the original units of measurement.

There is one special wrinkle that most modern statisticians use here: the "average" of the squared deviations is found by dividing the total of those squared deviations by $(n-1)$ instead of n. We will say a few words about this in a moment or two.

The average measure of squared deviation calculated in this way is called the *sample variance*, and is indicated by s^2. While n is the *sample size*, the number $n-1$ is called the *number of degrees of freedom*.

$$\text{Variance} = \frac{\text{total sum of the squares of deviations of observations from the mean}}{\text{number of degrees of freedom}} \qquad (3.3.2)$$

$$s^2 = \frac{\sum_{i=1}^{n} (y - \bar{y})^2}{n - 1}$$

Example 3.3.2

Our sample data on weights now give their sample *variance* as follows:

y	$y - \bar{y}$	$(y - \bar{y})^2$
1	−2	4
4	1	1
6	3	9
1	−2	4
6	3	9
2	−1	1
1	−2	4
21	0 (check)	32

$$\bar{y} = \frac{21}{7} = 3.0; \qquad s^2 = \frac{32}{6} = 5.33$$

Intuitive Explanation of Degrees of Freedom (here, $n-1$). Throughout this book we have considered our observations as a sample (subset) from a large population. For example, the 180 upper-class students of Mecca Community College are a sample of the total upper-class student body. After the sample has been taken, our purpose is to use the sample information we obtain to infer something about the larger population. Our thinking goes like this:

1. We have available a very large population from which to draw our sample.

2. We decide to take a sample of size n. (The manner in which the sample is taken is important and will be discussed later in the book.)

3. We say, then, that we have n "degrees of freedom" in the sample. Each of the n observations represents one "degree of freedom."

4. Once we have taken the sample and obtained the observations, we decide to characterize the sample by calculating some statistics. Our idea is to use the sample statistics to infer about the larger population. The first sample statistic we calculate is the arithmetic mean ($\bar{y}$), which represents one degree of freedom (1 d.f.) and now we have $\bar{y}$ and, in essence, only $n-1$ d.f. (pieces of sample information) left over to be used to obtain additional independent statistics.

5. Now, if we use $\bar{y}$ in calculating a new statistic, any averaging we do ought to be done using $(n-1)$ d.f. Therefore when we calculate the variance s^2, using $\sum (y-\bar{y})^2$ in the numerator of the formula, any averaging we do should use $n-1$ as the divisor. Thus we define the sample variance as in (3.3.2):

$$s^2 = \frac{\sum (y-\bar{y})^2}{n-1}$$

The reader is entitled to ask "If the above is an *intuitive* explanation of degrees of freedom, what would a *scientific* explanation be?" That question should be ducked by author and reader alike in the beginnings of statistical learning. The answer is involved in mathematical theory and usage connected with random behavior of statistics like s^2 across the universe of all possible different samples. We shall say a few words about this in a later chapter, but any reasonably clear as well as precise explanation does indeed require a level of study that logically comes after this book. Try to trust us that the long-run behavior of s^2 is generally considered better than it would be with n in the denominator, and that your discomfort here will be rewarded by certain comfortable simplifications in procedures later on.

C. Standard Deviation (s)

The variance s^2 is in squared units. In our weight example, that means s^2 is in units of (pounds)^2—squared pounds! To convert our measure of variability back into the original units of measurement, we take the square root of s^2. This gives us s, which we call the *sample standard deviation*, where standard deviation equals the square root of the variance.

$$s = \sqrt{s^2}$$

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

(3.3.3)

Example 3.3.3

Our data on weights gave us (Example 3.3.2) the variance $s^2 = 5.33$. So now we have the sample *standard deviation:*

$$s = \sqrt{s^2} = \sqrt{5.33} = 2.31 \text{ (pounds)}$$

The standard deviation is one of the most important statistics used in practice. We shall be using it a lot throughout this text.

D. Coefficient of Variation (C.V.)

The coefficient of variation expresses the standard deviation as a percentage of the arithmetic mean.

$$\frac{\text{Coefficient of}}{\text{variation (percent)}} = \frac{\text{standard deviation}}{\text{arithmetic mean}} \times 100$$

(3.3.4)

$$\text{C.V. (percent)} = \frac{s}{\bar{y}} \times 100 \text{ (percent)}$$

Example 3.3.4

For our sample of weight data (Examples 3.3.1–3.3.3) we collect results to obtain:

$$\text{C.V. (percent)} = \frac{s}{\bar{y}} \times 100 = \frac{2.31}{3.00} \times 100 = 77\%$$

The coefficient of variation is useful in comparing the relative variability of different kinds of characteristics. For example, it can be used to compare the variability of county income level in a state with the variability of county population size. Here one is comparing two different classes of measurement, one in dollars and the other in numbers of people. The coefficient of variation puts both of these on the basis of variability as a percentage of the mean, thus getting an index that is free of the unit of measurement. To avoid ambiguity in meaning, the coefficient of variation is best limited to use for data that are always positive and measured on a ratio scale.

Review Illustration. *Example 3.3.5.*

Consider the data on income from the review illustration in Example 3.2.4:

y
2400
2750
2180
2320
1930
3650
2180
2000
2190

Calculate: (a) range, (b) variance, (c) standard deviation, and (d) coefficient of variation.

a. Range:

$$\text{Range} = \text{largest} - \text{smallest y} \quad [\text{Formula (3.3.1)}]$$
$$= 3650 - 1930$$
$$\text{Range} = \$1720$$

b. Variance:

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1} \quad \text{[Formula (3.3.2)]}$$

y	$y - \bar{y}$	$(y - \bar{y})^2$
2400	0	0
2750	+ 350	122,500
2180	− 220	48,400
2320	− 80	6,400
1930	− 470	220,900
3650	+1250	1,562,500
2180	− 220	48,400
2000	− 400	160,000
2190	− 210	44,100

$\sum y = 21,600 \qquad 0 \qquad 2,213,200$
$\bar{y} = 2,400$

$$s^2 = \frac{2,213,200}{9 - 1} = \frac{2,213,200}{8}$$

$$s^2 = 276,650$$

c. Standard deviation:

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}} \quad \text{[Formula (3.3.3)]}$$

Using the result from (b), we have

$$s = \sqrt{276,650} = 525.9753$$

$$s = \$525.98$$

d. Coefficient of variation:

$$\text{Coefficient of variation (percent)} = \frac{s}{\bar{y}} \times 100 \quad \text{[Formula (3.3.4)]}$$

$$= \frac{525.98}{2400} \times 100$$

$$\text{C.V. (\%)} = 21.92\%$$

3.4 SOME COMMENTS ON TERMINOLOGY AND COMPUTATION

Students often complain that mathematics is a tyranny of exotic words and mysterious calculations. The same can be said for statistics. But then the same can be said for any subject matter dealing with ideas needing very precise definition and capable of quantitative measurement. Exotic words can be very helpful when their meaning is clear, and even mysterious calculations are welcome if they cut down labor.

We have introduced a variety of specialized words: mean, median, mode, range, variance, standard deviation, and coefficient of variation. Each has a precisely specified definition, giving a clearly stated characteristic of a set of data. You will find these technical terms useful in summarizing data and in communicating results to other investigators who will be using the same language.

In this connection, we should like to introduce you to a very special term that gets a great deal of use in a wide variety of ways in statistical analysis. It is *sum of squares*. We say it is a special term because it really does not mean strictly what it says. It means sum of *squared deviations*, the deviations being measured between the sample observations and their arithmetic mean. Thus for observations $y_1, y_2, y_3, \ldots, y_n$, we define:

$$\text{Sum of squares for } y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (3.4.1)$$

In terms of (3.4.1), the sample variance s^2 can be defined as:

$$s^2 = \frac{\text{Sum of squares for } y}{n-1}$$

Calculations of various sums of squares play a large part in statistical analysis. While an analyst can always hope to have available a machine which computes sums of squares automatically, the hope is often not realized. Then the labor of calculating $\sum (y - \bar{y})^2$ becomes important. You may have already foreseen from the examples on weights and incomes that computing a sum of squares can be a big chore. In our examples, $\bar{y} = 3$ exactly, or $\bar{y} = 2400$ exactly, so that figuring the differences $y_i - \bar{y}$, squaring them, and adding are no big deals. But you can be sure that in general $\bar{y}$ will come out with decimal places more numerous than in the observations themselves. And then the deviation-square task gets tedious. For that reason, statisticians have worked on the formula with a bit of algebra to come up with some alternative expressions that avoid taking the individual differences:

$$\text{Sum of squares for } y = \sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = \sum y^2 - n\bar{y}^2 \quad (3.4.2)$$

In Example 3.3.2 we worked out $\sum (y - \bar{y})^2$ directly, obtaining 32. By (3.4.2) we have:

y	y^2
1	1
4	16
6	36
1	1
6	36
2	4
1	1
21	95

$$\bar{y} = \frac{21}{7} = 3.0,$$

$$\sum (y - \bar{y})^2 = 95 - \frac{(21)^2}{7} = 95 - \frac{441}{7}$$

$$= 95 - 63 = 32$$

or

$$\sum (y - \bar{y})^2 = 95 - 7(3)^2 = 95 - 7(9)$$

$$= 95 - 63 = 32$$

$$s^2 = \frac{32}{6} = 5.33$$

In the review illustration of Example 3.3.5, we have $\sum (y - \bar{y})^2 = 2,213,200$. Use of (3.4.2) gives:

y	y^2
2,400	5,760,000
2,750	7,562,500
2,180	4,752,400
2,320	5,382,400
1,930	3,724,900
3,650	13,322,500
2,180	4,752,400
2,000	4,000,000
2,190	4,796,100
21,600	54,053,200

$$\bar{y} = \frac{21,600}{9} = 2400$$

$$\sum (y - \bar{y})^2 = 54,053,200 - \frac{(21,600)^2}{9}$$

$$= 54,053,200 - \frac{466,560,000}{9}$$

$$= 54,053,200 - 51,840,000 = 2,213,200$$

or

$$\sum (y - \bar{y})^2 = 54,053,200 - 9(2400)^2$$

$$= 54,053,200 - 9(5,760,000)$$

$$= 54,053,200 - 51,840,000 = 2,213,200$$

$$s^2 = \frac{2,213,200}{8} = 276,650$$

An additional comment on computing is required for the case where the data are available to you *only* in the form of a frequency table. As an example, suppose that the following table is all the information you have about the observations in a sample of size 209.

Cholesterol Level (mg/100 ml)	Frequency
125–149	4
150–174	13
175–199	30
200–224	42
225–249	51
250–274	34
275–299	23
300–324	8
325–349	3
350–374	1
	209

As we pointed out when we discussed the compilation of such tables in Chapter 2, the individual identities of the observations have been lost. All we know is the *number* of observations in each of the indicated intervals, and the best we can do is to approximate the real distribution by assuming the observations in any interval to be evenly distributed across that interval. So far as taking the *sum* of the observations is concerned, this assumption is equivalent to having all the observations in an interval measured as if at the *midpoint* of the interval: evenly distributed values would balance out "overs" and "unders" around that midpoint. Calculations of mean and standard deviation then proceed by adding in batches, as shown below.

The interval midpoints are found by carefully setting up the exact numerical boundaries on the scale of intervals and then taking the mean of each interval's beginning and ending points.

Interval limits	Cholesterol level	Frequency (f)	Interval midpoint (y)	Contribution to Σy (fy)	Σy^2 (fy²)
124.5–149.5	125–149	4	137	548	75,076
149.5–174.5	150–174	13	162	2,106	341,172
174.5–199.5	175–199	30	187	5,610	1,049,070
199.5–224.5	200–224	42	212	8,904	1,887,648
224.5–249.5	225–249	51	237	12,087	2,864,619
249.5–274.5	250–274	34	262	8,908	2,333,896
274.5–299.5	275–299	23	287	6,601	1,894,487
299.5–324.5	300–324	8	312	2,496	778,752
324.5–349.5	325–349	3	337	1,011	340,707
349.5–374.5	350–374	1	362	362	131,044
		209		48,633	11,696,471

$$\bar{y} = \frac{\Sigma y}{n} = \frac{48,633}{209} = 232.69,$$

$$\Sigma (y - \bar{y})^2 = \Sigma y^2 - n\bar{y}^2 = 11,696,471 - 209(232.69)^2$$

$$= 11,696,471 - 209(54,144.6361)$$

$$= 11,696,471 - 11,316,229 = 380,242,$$

$$s^2 = \frac{380,242}{208} = 1828.1,$$

$$s = \sqrt{1828.1} = 42.8$$

To find the *median* in such a table, we look for a point on the y scale at which 50 percent of the observations have accumulated. The procedure is seen best in graphical form. For $n = 209$, the median is at a point where we have the hypothetical accumulation of $.50 \times 209 = 104.5$ observations:

(Frequency) →   (4)   (13)   (30)   (42)   (51)

124.5   149.5   174.5   199.5   224.5   249.5
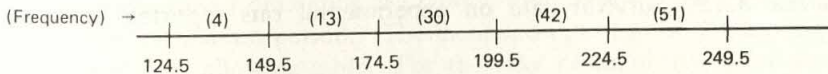
We have 89 observations up to 224.5 and so need $104.5 - 89 = 15.5$ observations out of the next interval. To get that many out of the 51 observations in the interval, we argue that the even distribution assumption tells us to go (15.5/51) of the way through the interval, and in distance that is (15.5/51) of the width of the interval. Thus

$$\text{Median} = 224.5 + \frac{15.5}{51} \ (25)$$

$$= 224.5 + \frac{387.5}{51}$$

$$= 224.5 + 7.6$$

$$= 232.1$$

Find the *range*, the *variance*, the *standard deviation*, and the *coefficient of variation* in each of the following cases on which you worked in the preceding set of exercises:

Exercise **3.4.1** "Misty" mouthwash data of Exercise 3.2.1.

Exercise **3.4.2** Chemical reaction times of Exercise 3.2.2.

Exercise **3.4.3** Detergent percentages in a soap product (Exercise 3.2.3).

Exercise **3.4.4** Liquid fabric finish used in a normal load of wash (Exercise 3.2.5).

Exercise **3.4.5** Pizza shell weights in two different production lines (Exercise 3.2.9).

Exercise **3.4.6** Monthly sales figures for a new product (Exercise 3.2.17).

Exercise **3.4.7** Diastolic blood pressure measurements (Exercise 3.2.20).

Exercise **3.4.8** Grouped data on height given in Exercise 3.2.21.

Exercise **3.4.9** Survival data on experimental rats reported in Exercise 3.2.22.

3.5 SOME COMMENTS ON LOOKING SUMMARY STATISTICS IN THE EYE

While all the measures of centrality and variability that we have discussed are useful indicators of general tendency and the amount of variation around it, the user must be very careful in applying such measures to real data. DON'T FORGET TO LOOK AT THE RAW DATA. This may seem a simple requirement to state, *but*, unfortunately, there are many cases where it is not possible. For example, when the data set is very large (say $n > 100$), looking at the raw data tables is not very helpful. Here, however, utilizing some of the tabular and graphical techniques already discussed *is* very helpful and should always be done. In other cases, only secondary data sources are available, such as grouped frequency tables published in journals. The electronic computer is also a source of some irritation on this matter; in many computer programs the original data never see the light of day! The printout shows only an analysis of the data.

Insist on seeing the actual data whenever you can. In the first place, the number of *faulty* data points discovered in many listings is usually quite surprising to the beginning data analyst. Error in recording a measurement, error in making the measurement, omission or duplication of a measurement, all of these can occur. They can sometimes be spotted by studying the raw data, sometimes not. Their existence is a constant hazard since all the statistics in the world will not correct these kinds of mistake.

So far as summary statistics are concerned, we have seen that each statistic that can be used has its own peculiar applicability. The only way we can check up on this is to see the data and try to reach a conclusion concerning the validity of using the statistic.

To further illustrate these points, let's consider the following real data on a botulism (a special kind of food poisoning) epidemic in La Plata, Argentina, in June 1957. The time from eating the tainted food until the onset of symptoms is called "the incubation period." The incubation periods of the first 21 cases are shown in the following table. (For the sake of simplicity, three numbers have been modified by one hour so that the arithmetic is easy; this does not change the conclusions.)

Incubation Periods in Botulism Victims (hours)		
1. 14	8. 48	15. 36
2. 19	9. 45	16. 43
3. 20	10. 48	17. 29
4. 20	11. 21	18. 19
5. 32	12. 78	19. 43
6. 34	13. 17	20. 85
7. 36	14. 20	21. 91

These data were collected to get some idea about the characteristics of the incubation period of the disease in this particular outbreak. The following points should be determined:

(a) average incubation period for botulism in this epidemic and (b) degree of variability of incubation period.

The immediate response would be to calculate the arithmetic mean to determine point (a) and to calculate the standard deviation for point (b). In this case

$$\bar{y} = 38 \text{ hours}$$

$$s = 22.4 \text{ hours} \quad \text{(coefficient of variation} = 58.9\%)$$

Verify these calculations.

However, this is an inadequate answer. Notice that the standard deviation is almost 60 percent as large as the mean. Whenever the sample standard deviation is this large, one should examine the data carefully. Quite often one finds a few very large or very small observations. These should be investigated to make sure that they are valid and belong. Table 3.5.1 shows all of the individual contributions to the standard deviation.

TABLE 3.5.1 Standard Deviation Composition

Case Number	y	$y - \bar{y}$	$(y - \bar{y})^2$
1	14	−24	576
2	19	−19	361
3	20	−18	324
4	20	−18	324
5	32	− 6	36
6	34	− 4	16
7	36	− 2	4
8	48	+10	100
9	45	+ 7	49
10	48	+10	100
11	21	−17	289
12	78	+40	1,600
13	17	−21	441
14	20	−18	324
15	36	− 2	4
16	43	+ 5	25
17	29	− 9	81
18	19	−19	361
19	43	+ 5	25
20	85	+47	2,209
21	91	+53	2,809
Totals	798	0 (check)	10,058

$$\bar{y} = \frac{798}{21} = 38; \qquad s = \sqrt{\frac{10,058}{20}} = \sqrt{502.9} = 22.4$$

$$[\text{Median} = y_{(11)} = y_6 = 34]$$

Notice that the incubation periods of three of the cases (12, 20, 21; namely 78, 85, and 91 hours) account for

$$\frac{1600 + 2209 + 2809}{10,058} \times 100\% = 65.8\%$$

of the total sum of squares used in calculating the standard deviation. These three extremely long incubation periods are the major contributors to the large standard deviation. Furthermore since they are all on one side of the arithmetic mean, we can also say the sample distribution is *skewed* (*to the right*, since the

TABLE 3.5.2 Distribution of
Incubation Periods

Class Interval	Frequency
10–24	8
25–39	5
40–54	5
55–69	0
70–84	1
85–99	2
Total	21

distribution has a stretched-out right side). By plotting the data, using a few intervals, say six, we can see this in Table 3.5.2.
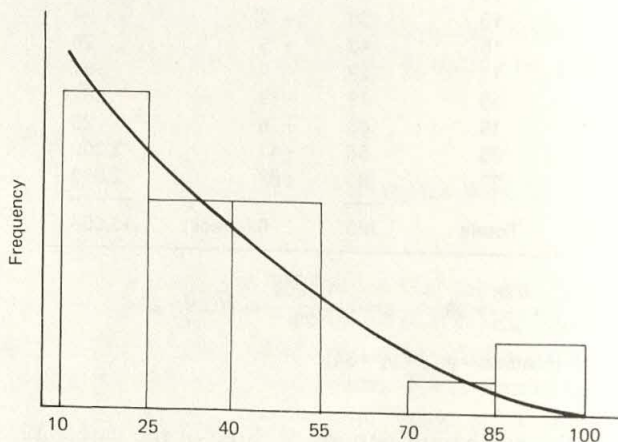
As a result of the histogram plot, we might go further and conjecture the shape of the incubation period distribution for a theoretical botulism epidemic to look like the smooth curve put through the data. While we cannot prove this, at least it is the possible foundation for a hypothesis about the shape of the incubation period for a typical botulism epidemic.

Thus by carefully examining the individual data and the elements used in calculating s, and then using plotting of grouped data, we have arrived at a reasonable hypothesis or conjecture for the distribution of the incubation period of botulism.

Another guideline useful for detecting the unusual behavior of some of the observations is to determine by how many sample standard-deviation units each observation deviates from the sample mean. For example, the botulism data show the following (Table 3.5.3):

TABLE 3.5.3 Standardized Deviations of Incubation Periods

Incubation Period (Hours) y	Deviation from $\bar{y}$ $y - \bar{y}$	Standardized deviation $\dfrac{y - \bar{y}}{s}$
14	−24	−1.07
19	−19	− .85
20	−18	− .80
20	−18	− .80
32.	− 6	− .27
34	− 4	− .18
36	− 2	− .09
48	+10	+ .45
45	+ 7	+ .31
48	+10	+ .45
21	−17	− .76
78	+40	+1.78
17	−21	− .94
20	−18	− .80
36	− 2	− .09
43	+ 5	+ .22
29	− 9	− .40
19	−19	− .85
43	+ 5	+ .22
85	+47	+2.10
91	+53	+2.36

$\bar{y} = 38, \quad s = 22.4$

As you can see, 17 of the 21 observations are within one standard deviation of the sample mean. Furthermore, three of the four that deviate by more than one standard deviation are on the positive side. This, plus the fact that 13 of the 21 are small negative deviations, leads to the conjecture that the data are skewed to the right; that is, they have an asymmetrical sample distribution with right-hand stretchout.

SUMMARY EXERCISES

3.5.1 The following table gives data on per capita income in 134 counties reported in a study of food programs in social service.

		Per Capita Income ($) in 134 Counties Having Food Distribution or Food-Stamp Programs			
2400	1820	3280	2550	1770	2850
2210	5400	1870	2660	1800	1200
2150	2175	2300	1710	1600	2620
2750	2780	1680	2320	2310	2000
1975	1800	1835	700	3450	1760
2650	3300	1400	478	2210	1695
3570	1625	530	1830	2460	2350
2020	1910	1720	3150	1830	1625
2710	2200	2010	1840	810	1900
2180	1800	1920	1815	2510	520
1970	1750	1675	1250	1930	2610
3850	2700	800	1950	1550	2480
2950	2230	450	550	3650	3100
1700	2560	2450	2800	1960	2015
1730	2220	1675	1650	400	2330
1760	2540	1695	2100	1810	1575
515	1770	2760	1710	750	1860
3350	500	3050	3200	3250	3900
1850	2870	600	1750	2470	1990
2600	2515	1690	2790	2180	1740
1940	1300	3700	2770	2900	3975
1910	1000	900	1775	1925	2190
2860					482

a. Describe the information contained in these data by constructing a frequency table (take intervals 250–749.99, 750–1249.99, 1250–1749.99, etc., and round the midpoints to 500, 1000, 1500, etc.). Using this table, draw a frequency histogram and a frequency polygon.

b. Calculate the following statistics: (i) range, (ii) mean, (iii) median, (iv) standard deviation, and (v) coefficient of variation.

3.5.2 To get some practice in dealing with data that relate to a variety of characteristics in a single study, let us consider again the survey of upper-class students in Mecca Community College (Table 1.2.1). To keep down computational labor, and to see how different samples behave, separate the 180 observations into four samples:

Sample Number	Student Numbers
1	1–45
2	46–90
3	91–135
4	136–180

Each of the four samples has size $n = 45$. We shall assume that the method of putting together the survey of 180 students allows us to consider each of the samples of size 45 to be a valid random sample for representing the upper-class students.

By instructor's edict, town-hall meeting of the class, drawing numbers out of a hat, or any other nonviolent procedure, distribute the four samples around and around, one to each student. Let each student then work on his or her sample, using the data to give answers to the following questions. Extra credit is available for coming up with additional features that can be wrung out of the data.

a. What percentage of the sample is: (i) male and (ii) female?

b. What is the "average" commuting distance of the students in the sample? (Look at mean and median, distinguish between them.) What is the standard deviation of the commuting distance? The coefficient of variation? What does the distribution of commuting distance look like? Is there a difference in commuting distance between that for females and that for males?

c. What can we say about political party preference among the students in the sample? Does it differ with sex? Does it change depending on how far one commutes?

d. How do these students vote on the "legalizing of marijuana" question? What percentage of the students agree to legalization? How does this percentage differ: (i) between the two sexes and (ii) among the four categories of political party preference?

e. What does the sample distribution of the GPA's look like? What are its mean and standard deviation? How does the G.P.A. differ between the sexes, if at all? Is there a relation between G.P.A. and commuting distance?

f. If you were to use the answers to the above questions on the sample of 45 to infer conclusions about the sample of 180, on which of your answers would you feel you couldn't be far wrong?

4.1 DESCRIPTION AND INFERENCE IN STATISTICS

Up to now we have been considering orderly procedures for describing and summarizing a collection of observed numerical data. In most cases of practical interest, such a collection of data is only a small fraction of all possible observations that could in principle be made. It is a *sample* from a *universe*, or *population*, of outcomes in the particular process under study.

The data on community-college upper-class students in Chapter 1 came from one sample of upper-class students in one specific year in one specific college system. A different sample would yield somewhat different data for the upper-class population of that system in that year. A different year would give a different set of data from all possible sets of data *over time* in the specific system. A different system would yield a different collection of data on the population of community-college upper-class students *in a wider region*.

The data in Exercise 2.3.3 for the source of funds in the XYZ Health Department in 1960 and 1965 are specific for that health department in the stated years. Different health departments and/or different years would give different data on funding sources in health departments.

The age distribution reported on page 44 for patients treated for carcinoma of the cervix in the Barnard Free Skin and Cancer Hospital, St. Louis, during the period

4

Statistics and Chance

4

1906–1926 is precise for that specific hospital during that specific span of years. Observation of ages of such patients will give different data for other possible hospitals and other observable years.

While the *descriptive statistics* that we have been studying are extremely valuable for assessing what we have observed, it is customary for the data to be given the additional obligation of telling us something dependable about the *universe* from which our observations were drawn as a sample. What can we say about upper-class students at Mecca Community College *in the next 5 years*? What can we say about upper-class students in American community colleges generally? What can be said about Health Department funding across the state of North Carolina, now and in 1980 and 1985? To what extent do the Barnard Hospital data represent the age distribution of all Americans now having carcinoma of the cervix?

In the majority of cases, the purpose of collecting data in the first place is to form the basis of making a generalization to a much larger universe—the total population in a country, the underlying biological process, the ongoing production line, or the actual natural law, from which the observations are taken as a sample. This is the field of *statistical inference*; from a particular sample we

want to be able to *infer* one or another characteristic of the universe that produced the sample. The report in Exercise 2.3.2 concerning the employment of persons 14 years of age and over as of March 1968 came from such an inference. The Bureau of the Census did not count and describe every employed person in the United States in March 1968; it used the data of a sample scientifically chosen by the Bureau of Labor Statistics and generalized in a dependable manner to the entire country.

The logic that we use in statistical inference is *inductive* logic, arguing from particular cases to the general law. Such inductive argument has two main aspects: (a) *estimation* of critical characteristics of the universe and (b) *test of a hypothesis* concerning the universe. The latter aspect is a fundamental feature of the time-honored "scientific method": form a hypothesis about a state of nature, make observations on that state, cor..pare observations with hypothesis, and accept or reject the hypothesis according as the observations are or are not consistent with the hypothesis.

Estimation is also an essential feature of scientific investigation, for it is on the basis of observations that the scientist arrives at values of critical physical constants whenever they can not be found by purely theoretical logic. Until more advanced mathematical methods were applied, the value of π was an estimate, for example, 3 in the Bible, $(16/9)^2$ in ancient Egyptian writings. While the so-called *absolute* constants, like π, can eventually be determined mathematically to within any desired degree of accuracy, there remains the vast variety of processes in nature wherein the constants of importance can be approached *only* through observation—the growth rate of a bacterium species, the constants defining the accident rate on a particular superhighway, the proportion of recoveries from a given disease after use of a given drug, and so on and on.

The essential features of a situation that requires statistical inference are: (a) *variability* in the population—the circumstance that not all elements of the universe exhibit the same value for the characteristic under study and (b) *sampling*—the circumstance that our observations constitute only a fraction of all possible observations in the universe.

Variability in the population causes it to exhibit what is called *random behavior*. Different members of the population have different values of the characteristic under study, and we cannot know ahead of time the exact value that will be shown by a member chosen "at random" from the population.

When we have chosen "at random" a number of members of the population, thus forming a sample, and have observed their values, we have seen the outcome of one experiment on the population's random behavior. We know that this cannot give us the precise truth about the entire population. What we should reasonably want, therefore, is an inductive procedure that will enable us to have a definable level of confidence that the truth is within specifiable

practical limits, or that a decision we make about a hypothesis will be correct.

Terms like "random," "at random," and "level of confidence" carry certain intuitive meanings, but they must be made precise in some quantitative structure that we can apply mathematically. Notions of chance occurrences, the likelihood of an event, and betting odds have been in man's thinking (and acting!) from the earliest times. Quantifying the notions in a coherent framework dates from the middle of the seventeenth century. Since that time mathematicians and philosophers have continued the building and refinement of the structure. The mathematical subject matter is called *probability* (or *mathematical probability* or *probability theory*). In this chapter we shall try to give a grasp of those elements of the subject most needed for statistical inference.

4.2 DEFINITION OF PROBABILITY

DEFINITION 4.2.1 At least intuitively, if we wish to assign a numerical value to the likelihood of a chance event, we would use as our number the proportion of times the event occurs in an enormously long sequence of opportunities for the event to happen.

We think of 1/2 as a reasonable measure of the likelihood of getting "head" when we toss a coin because we believe that a balanced coin should turn up "head" one half of the time if the coin is tossed a very great number of times. If we roll a balanced six-sided die whose faces are marked with the customary number of dots (1, 2, 3, 4, 5, 6), we would consider 1/6 as a reasonable measure of the likelihood of having the die show four dots on its top face after coming to rest, because we believe that in a very long sequence of rolls the face showing "4" should end up in top position one-sixth of the time.

On such basis we could agree to start the quantifying of likelihood by deciding that the numerical measure of the likelihood of an event shall be a number between 0 and 1. This has been the established convention from the earliest days of the mathematical development of the subject. The next basic convention is to give the name *probability* to the numerical measure of likelihood.

Thus we say that the probability of tossing a head is 1/2, the probability of rolling a "4" with a single die is 1/6; and we use the notation

$$P(\text{head}) = 1/2, \qquad P(\text{four}) = 1/6.$$

If there is any danger of confusion as to the chance situation being discussed, we can insert into the parenthesized expression a statement of the underlying condition, using a vertical bar to set it off:

$$P(\text{head} \mid \text{toss of a balanced coin}) = 1/2,$$

$$P(\text{four} \mid \text{roll of a balanced die}) = 1/6.$$

The vertical bar is often read "given," so that we read the above statements as "the probability of *head*, given the toss of a balanced coin, is 1/2," "the probability of *four*, given the roll of a balanced die, is 1/6." We regularly omit such statements of condition when the chance situation has been clearly identified at the beginning of the discussion into which the probability enters.

The foundations of a mathematical theory of probability were laid by the great philosopher–mathematician Blaise Pascal (1623–1662) and the eminent mathematician Pierre de Fermat (1601–1665), in an exchange of correspondence initiated by questions to Pascal by a French nobleman who found contradictions between standard gambling rules and his own experience. From this early work came the original—now called the *classical*—definition of probability:

DEFINITION 4.2.2 If in a single trial of a chance situation there are *t* different possible fundamental outcomes that are exhaustive, mutually exclusive, and equally likely, and if *f* of these outcomes are favorable to an event A, then the mathematical probability of A is defined as the ratio *f/t*.

Let us illustrate the classical definition of probability with five examples that help us to understand words and concepts like *exhaustive, mutually exclusive,* and *equally likely* in the definition.

Example 4.2.1.   Toss of a balanced penny.

Here there are two possible outcomes, *head* or *tail*. These exhaust the possible outcomes. The two outcomes are mutually exclusive; that is, if a head occurs, a tail cannot, or if a tail occurs, head cannot. Finally, since the penny is balanced, there is an equal chance of getting a head or a tail in the toss of the penny. Hence $t = 2$ in the above definition.

Example 4.2.2.   Toss of a balanced coin three times.

Here the fundamental outcomes can be enumerated as the sequences HHH, HHT, HTH, HTT, THH, THT, TTH, and TTT. Thus $t = 8$. If A is the event *two heads in the three tosses*, then the fundamental outcomes favorable to A are HHT, HTH, and THH. Hence $f = 3$ and $P(A) = 3/8$.

Example 4.2.3.   Roll of a balanced die.

Here the exhaustive, mutually exclusive, and equally likely fundamental outcomes are the six different numbers of dots that can appear face up after the die comes to rest: 1, 2, 3, 4, 5, 6. If A is the event *four*, then only the outcome 4 is favorable, and so $P(A) = 1/6$. If B is the event *a number divisible by 3*, then the favorable outcomes are 3 and 6, whence $f = 2$ and $P(B) = 2/6 = 1/3$.

Example 4.2.4.   Roll of a pair of balanced dice.

Here we can identify a fundamental outcome by giving a pair of numbers, the first referring to Die No. 1 and the second referring to Die No. 2. Thus there are 36 fundamental outcomes, shown as follows.

Fundamental Outcomes					
1,1	2,1	3,1	4,1	5,1	6,1
1,2	2,2	3,2	4,2	5,2	6,2
1,3	2,3	3,3	4,3	5,3	6,3
1,4	2,4	3,4	4,4	5,4	6,4
1,5	2,5	3,5	4,5	5,5	6,5
1,6	2,6	3,6	4,6	5,6	6,6

In this example $t = 36$. Suppose that we are interested in the total number of dots showing on the top faces of the two dice when they come to rest after being tossed. In dice games this total is usually called the *point* that is rolled. If A is the event *point 7*, then the fundamental outcomes favorable to A are $(1, 6)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 2)$, and $(6, 1)$. Thus $f = 6$, and $P(\text{point } 7) = 6/36 = 1/6$.

Example 4.2.5.   Drawing an upper-class student at random from the Mecca Community College sample.

To draw a unit "at random" from a collection of units means to choose a unit by some process whereby every unit in the collection has equal chance of being drawn. One elementary way of doing this would be to put into a box an appropriately labeled tag for each unit in the collection, shake the box vigorously so as to mix the tags well, and then draw one tag from the box. In such a random drawing, the individual units in the collection are the set of t exhaustive, mutually exclusive, and equally likely outcomes referred to in the classical definition of probability.

In the present example we thus have $t = 180$. Suppose we ask for the probability that a student drawn at random will have no party preference politically. We have only to count the number of Ns in the data set, find it to be 64, so that $f = 64$, and then argue:

$$P(\text{no party preference}) = \frac{64}{180} = \frac{16}{45} = 0.356.$$

In calculating probabilities by the classical definition, the whole problem resides in making an accurate count of the total number of possible outcomes and the number favorable to the event under consideration. This can become very complicated as soon as we move away from straightforward cases. Making such counts comes under the heading of "combinatorial" mathematics, a subject beyond our needs for this book. Also, most practical problems do not have such nice cleanly defined outcomes that are equally likely. Thus the classical definition of probability had to be generalized in two directions: (a) to allow the fundamental outcomes to differ as to likelihood and (b) to cover cases where the possible outcomes are too numerous to count. Before taking up this aspect of probability, let's consider the practical implications of a well-defined classical case.

4.3 THE PRACTICAL MEANING OF PROBABILITY

Let us return to the matter of interpretation with which we began. The probability $P(A)$ of an event is the long-run proportion of times that the event A occurs in a sequence of trials of the experiment. The proportion of A occurrences in any finite number n of trials keeps changing as n changes; $P(A)$ is the limiting value of these proportions as n increases without bound, going on to infinity.

For example, consider the experiment of rolling a pair of dice and getting a total of 7. We have seen that the probability of getting a total of 7 is 1/6 ($= 0.167$). In two different actual experiments of rolling a pair of dice 504 times, the first 18 rolls yielded the cumulative records shown in Table 4.3.1. Notice the wide oscillation of the proportion of 7s rolled in these early stages of the sequence of trials. The cumulative records for trials 15 through 504 are shown graphically in the diagrams of Figure 4.3.1 on the next page.

TABLE 4.3.1

	Experiment 1			Experiment 2	
Number of rolls	Cumulative number of 7s	Cumulative proportion of 7s	Number of rolls	Cumulative number of 7s	Cumulative proportion of 7s
1	1[a]	1.000	1	0	.000
2	1	.500	2	0	.000
3	1	.333	3	0	.000
4	1	.250	4	0	.000
5	1	.200	5	0	.000
6	1	.167	6	1	.167
7	1	.143	7	1	.143
8	1	.125	8	2	.250
9	1	.111	9	2	.222
10	1	.100	10	2	.200
11	1	.091	11	2	.182
12	1	.083	12	2	.167
13	2[a]	.154	13	3	.231
14	2	.143	14	3	.214
15	3[a]	.200	15	3	.200
16	3	.188	16	3	.188
17	4[a]	.235	17	3	.176
18	4	.222	18	3	.167

[a].On the first roll, a 7 was obtained, and the second 7 was gotten on the 13th roll, thus making a cumulative total of two rolls each getting a 7; the third 7 obtained on the 15th roll, and so on.
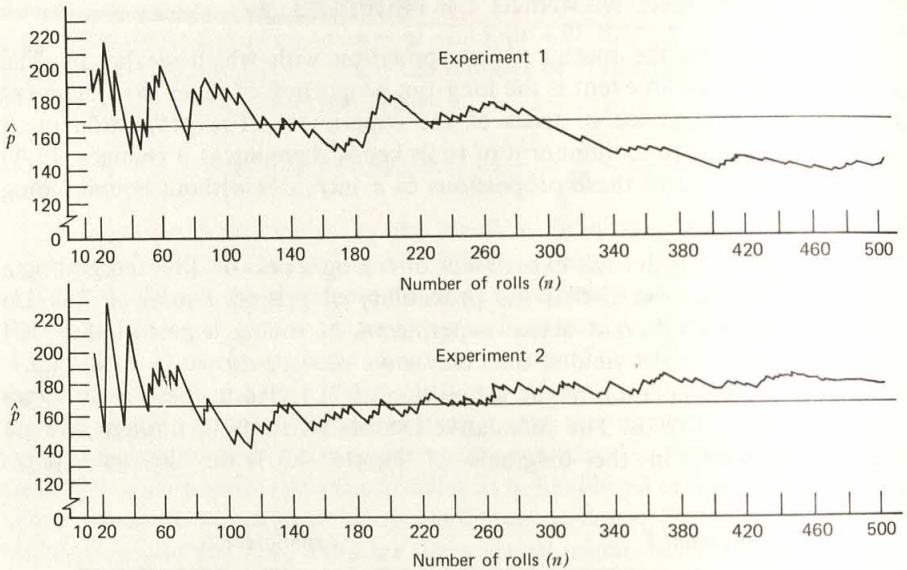
FIGURE 4.3.1 Proportion ($\hat{p}$) of n pair-of-dice rolls giving point 7.

In each of the two experimental sequences there is evidence that the proportion of 7s is making some effort to settle down toward 1/6. However, 504 rolls are not nearly long enough a "long run" to bring the proportion consistently very close to 1/6. We can calculate later in the chapter that something over 5000 rolls are required to give 95 percent probability that the observed proportion will be within .01 of 1/6. To raise that probability to 99.99 percent requires 21,000 rolls. Three such sequences performed by an eager electronic computer had representative stages in their history as shown in Table 4.3.2. Thus the above indicates that it takes a very large number of rolls of a pair of dice to arrive at the expected probability level.

TABLE 4.3.2

Number of rolls	Cumulative Proportion of 7s		
	Sequence 1	Sequence 2	Sequence 3
100	.1700	.1900	.1400
200	.1900	.1500	.1450
300	.1800	.1367	.1500
400	.1775	.1300	.1500
500	.1720	.1320	.1500
600	.1667	.1250	.1533
700	.1629	.1286	.1514
800	.1675	.1275	.1513
900	.1689	.1267	.1467
1000	.1700	.1370	.1450
1500	.1633	.1453	.1453
2000	.1620	.1600	.1485
2500	.1660	.1568	.1496
3000	.1633	.1620	.1517
3500	.1614	.1669	.1523
4000	.1628	.1675	.1565
4500	.1676	.1698	.1549
5000	.1660	.1692	.1576
6000	.1677	.1647	.1575
7000	.1650	.1647	.1583
8000	.1674	.1646	.1599
9000	.1676	.1660	.1599
10000	.1664	.1658	.1613
12500	.1637	.1680	.1639
15000	.1629	.1691	.1655
17500	.1622	.1694	.1655
20000	.1628	.1681	.1647
21000	.1624	.1690	.1643

In other practical situations the argument deals with the prevalence of a certain characteristic in a finite population of elements: lung cancer in a population of male smokers, individual weight between 1 and 2 pounds in a school of fish, tree girth between 6 and 12 inches in a forest, favorable opinion about a referendum question in a community of voters. If we designate by A the characteristic of interest, then $P(A)$ is the measure of likelihood that A will be exhibited by a single individual member drawn at random from the population. With respect to the entire population, $P(A)$ has the important related meaning as the proportion of the population that is composed of A-type elements. Thus $P(\text{lung cancer}) = .05$ indicates that 5 percent of the

related population has lung cancer; P(fish weight between 1 and 2 pounds) = .70 indicates that 70 percent of the fish in the associated population are of weight 1–2 pounds each. In a matter of opinion in a population of 5000 voters, P(favorable opinion) = 3/10 can tell us that 30 percent of the voters in that population hold a favorable opinion.

The above interpretations of the probability of an event, in terms of the limit of the proportion of times the event occurs in an unending sequence of trials, or the proportion of an entire population exhibiting the characteristic which is specified by the event, give practical meaning for long-run or overall experience. But what of the practical meaning for a *single* trial? For example, we know that the probability of rolling point 7 in a single fair roll of a pair of dice is 1/6, but when that roll is actually performed the result will be either a 7 or not a 7. All that probability tells us is that the latter outcome is five times as likely as the former, since $P(7) = 1/6$ while P(not a 7) $= 1 - (1/6) = (5/6)$.*

The gambler uses probability to set fair "odds" for a single trial. He argues, for example, that since not-7 is five times as likely as 7, the person who bets $1 that he will roll a 7 should receive $5 if he actually rolls it, and the statement is made that "fair odds are 5-to-1 (or 5:1)." These odds are considered "fair" on the basis of equating "expected" gain with "expected" loss, expectation being the result of weighting the gain (or loss) by the probability of attaining it. If the roll is made at 5:1 odds, the expected gain is $(1/6)(\$5) = \$(5/6)$, and the expected loss is $(5/6)(\$1) = \$(5/6)$. Such expectations are again "long-run" concepts since they involve probabilities. Thus the logic of setting such fair odds is again a "long-run" argument: in the long run the gambler will have gains to balance losses.

DEFINITION 4.3.1 In general the mathematical odds in favor of A (against not-A) are $a:b$, where a and b are integers satisfying

$$\frac{a}{b} = \frac{P(A)}{1 - P(A)}.$$

---

* It must be a general rule that for any event A, $P(A) + P(\text{not-}A) = 1$, since the events A and not-A exhaust all possibilities, thus accounting for the totality of probability while at the same time having no overlap of outcomes common to them both.

Thus the mathematical odds in favor of rolling 7 are given by

$$\frac{\frac{1}{6}}{1-\frac{1}{6}} = \frac{1}{6-1} = \frac{1}{5},$$

so that we say the mathematical odds are $1:5$ in favor of rolling 7. In cases where $a < b$ it is often preferred to state the odds as $b:a$ *against A*. Thus the odds for rolling 7 can be stated as $5:1$ against rolling 7.

Gambling odds are always taken as $b:a$ since these odds are intended to balance out expectations based on mathematical odds, and

$$\frac{a}{b} \cdot \frac{b}{a} = 1 = \frac{1}{1} = 1:1.$$

Thus in summary we match reality to the formal definition of mathematical probability by interpreting $P(A)$, the probability of event A, as: (a) the long-run proportion of times that A occurs in an unending sequence of trials, (b) the proportion of the population that is composed of A-type elements, (c) the single-trial likelihood of the occurrence of A, expressible as $P(A):$ $[1 - P(A)]$ odds in favor of the occurrence of A.

EXERCISES

4.3.1 In Example 4.2.2 we calculated the probability of getting exactly two heads in three tosses of a balanced coin. We found $P(2 \text{ heads} | 3 \text{ coin tosses}) = 3/8$. Calculate the probabilities of the other possible events; that is, find $P(0 \text{ heads})$, $P(1 \text{ head})$, $P(3 \text{ heads})$.

4.3.2 Make the probability calculations for all possible numbers of heads in *four* tosses of a balanced coin.

4.3.3 In Example 4.2.4 we considered the rolling of a pair of fair dice and calculated as 1/6 the probability of rolling "point" 7. Calculate the probability of each possible "point" that can be rolled. That is, find $P(\text{point } 2)$, $P(\text{point } 3), \ldots, P(\text{point } 11)$, $P(\text{point } 12)$.

4.3.4 A deck of bridge playing cards contains 52 cards, composed of 13 "denominations" of each of four "suits." The suits are *clubs, diamonds, hearts,* and *spades.* Clubs and spades are black in color, and diamonds and hearts are red. The denominations are 2, 3, 4, 5, 6, 7, 8, 9, 10, jack, queen, king, and ace. Such a deck is well shuffled and a card is drawn at random. What is the probability that the card will be: (a) a heart, (b) red card, (c) ace, (d) face card (jack, queen, or king), (e) card with denomination under 10, (f) one-eyed jack (jack of hearts and jack of spades are pictured in profile, showing one eye)?

4.3.5 If a student is drawn at random from the Mecca Community College sample, what is the probability that the student: (a) is a female, (b) has preference for the Republican party, (c) has no opinion about legalizing marijuana, (d) commutes more than 20 miles?

4.3.6 If the student drawn at random in Exercise 4.3.5 is seen to be a female, then what are the probabilities in (b), (c), and (d)?

4.3.7 What is the probability that a randomly chosen Mecca College sample male will disagree on legalizing marijuana? What is the corresponding probability for a female?

4.3.8 What is the probability that a sample student randomly chosen from those commuting less than 5 miles will have Democrat political preference? What is the corresponding probability for a student commuting more than 20 miles?

4.3.9 A cage of inoculated laboratory mice contains four males and four females. An adjoining cage of noninoculated mice has 10 males and six females. By accident the barrier between the cages is released, and in the morning the experimenter finds the mice all mixed up in a single cage. If one mouse is selected at random, what is the probability that it is: (a) inoculated, (b) male, (c) inoculated, given that it is male?

4.3.10 Eighteen students went on an all-day hike. Six got sunburned, five got bitten by chiggers, nine made it without either of those misfortunes. What is the probability that: (a) a sunburned hiker escaped the chiggers and (b) a bitten hiker was also sunburned? (Hint: use logic on the counts to identify how many hikers were both bitten and burned.)

4.3.11 In one of the manufacturing plants of a phonograph record company, machines A, B, and C turn out pressings, machine A producing 30 percent of the total output, machine B 50 percent, and machine C 20 percent. Each machine has a certain fraction defective in its output: A, 3 percent; B, 2 percent; C, 4 percent. Out of the joint production a record is drawn at random and found to be defective. What are the probabilities that it was pressed by machine A, B, or C, respectively? (Hint: taking total production as the general amount T will enable you to make all of the necessary calculations.)

4.4 INDEPENDENT EVENTS

Of great importance in the analysis of chance situations is the notion of *independence in the probability sense.* An intuitive idea is that events A and B are independent (in the probability sense) if the occurrence of one of them has no effect on the probability of the occurrence of the other. We apply such an idea naturally when A and B are events in successive tosses of a coin or in successive rolls of a pair of dice. It is the basis of the gambler's rule that "a coin has neither memory nor conscience."

The development of the mathematical definition of independence usually proceeds through consideration of *conditional probability*—the probability of an event *conditional on* the occurrence of another event. We do not wish to go into this subject in any detail, and so ask the reader to accept a *definition* of independence and settle for some numerical examples of the implications respecting conditional probability.

The *joint* occurrence of events A and B is called the *compound* (or *intersection*) *event AB.* Some examples are: head on two successive tosses of a coin, "4" and then "7" in two successive rolls of a pair of dice, male and then female in two random selections of students from the community-college sample, male *and* democratic party preference in a random selection of *one* student from the college sample, sunburned and bitten by chiggers in Exercise 4.3.10.

> **DEFINITION 4.4.1** The events A and B are *independent* if and only if
> $$P(AB) = P(A) \cdot P(B).$$

It is because of such independence that

$$P(HH \mid \text{coin tosses}) = P(H) \cdot P(H) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$

$$P(\text{"4", "7"} \mid \text{dice rolls}) = P(\text{"4"}) \cdot P(\text{"7"}) = \frac{1}{12} \cdot \frac{1}{6} = \frac{1}{72}.$$

Tallying the Mecca College sample data leads to the following table of the numbers of students distributed as to sex and political party preference:

Political party pref. Sex	Democrat (D)	Republican (R)	Other (O)	No. pref. (N)	Total
Female (F)	27	21	1	27	76
Male (M)	33	29	5	37	104
Total	60	50	6	64	180

In the data of the community-college sample above, there are 104 males and 76 females. Hence for a single random drawing

$$P(M) = \frac{104}{180} = \frac{26}{45}, \qquad P(F) = \frac{76}{180} = \frac{19}{45}.$$

If we make two random selections of students by drawing a name, noting it, replacing it in the collection, then making a second random selection, we have an example of independent events, since

$$P[(M, F) \mid \text{drawing with replacement}] = \frac{104}{180} \cdot \frac{76}{180} = P(M) \cdot P(F)$$

But if we do not replace the first draw before making the second selection, then with only 179 names left and 76 of them are female, the joint probability is written:

$$P[(\text{male, female}) \mid \text{drawing without replacement}] = \frac{104}{180} \cdot \frac{76}{179},$$

and this is *not* P(male) · P(female). Hence such draws are not independent.

The event "male *and* Democratic party preference" is the outcome of the random drawing of one student provided that student is one of the 33 students shown in the corresponding cell of the table. Hence

$$P(MD) = \frac{33}{180} = \frac{11}{60}.$$

But

$$P(M) = \frac{104}{180} = \frac{26}{45}, \qquad P(D) = \frac{60}{180} = \frac{1}{3},$$

so that

$$P(M) \cdot P(D) = \frac{26}{45} \cdot \frac{1}{3} = \frac{26}{135}.$$

Thus $P(MD) \neq P(M) \cdot P(D)$, so that *male* and *Democrat* are not independent.

In fact, we can say that *sex* and *political party preference* are not independent, since independence of those characteristics would require that the multiplication equality hold for all eight sex–party combinations, whereas we have already seen the equality fail for the Male–Democrat combination.

More than two events are considered independent if multiplication equalities hold for all compound events formed of any two of them, any three of them, any four of them, and so on.

DEFINITION 4.4.2 The event $A_1, \ldots, A_k$ are *independent* events if and only if the probability of the joint occurrence of any collection of the individual events equals the product of the probabilities of the individual events in the collection.

In the practical applications of probability theory, independence is of special importance when an investigator makes repeated "trials" of a chance situation, such as repeated tosses of a coin, repeated rolls of a pair of dice, a treatment applied to a number of patients, blood pressure measured on a number of men, or repeated blood-pressure measurements made on a single individual. Such repeated trials are called *independent repeated trials* if the same probability structure applies to all trials *and* the equalities in Definition 4.4.2 hold for the events made up of the outcome on trial 1, the outcome on trial 2, and so on through the outcome on trial *n*. The most commonly used methods of statistical inference require that the observed data come from independent repeated trials. Hence in any experiment to which such statistical inference is to be applied, an important consideration is assuring that the conditions for independent repeated trials are satisfied. Thus the restrictions "fair" coin and "fair" toss, "fair" dice and "fair" roll, and the requirements that we shall meet in Chapter 5 for having the proper kind of sample for generalizing to a population.

In independent repeated tosses of a fair coin, for example,

$$P(5 \text{ successive heads}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{32} = .03125;$$

and in independent repeated rolls of a pair of fair dice,

$$P(5 \text{ successive 7s}) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{7776} = .0001286.$$

(Looking at these probability values, the reader is likely to say that if *he* was present at the occurrence of five successive heads or five successive 7s in any short-run experience, *he* would doubt the validity of the assumption "fair." At that moment he will have grasped the logic underlying one major part of statistical inference!).

4.5 BERNOULLI TRIALS. THE BINOMIAL DISTRIBUTION

A particularly interesting class of experiments involving n independent repeated trials is that in which each trial has just two possible outcomes. The coin-tossing experiment is an example. Each item drawn from an industrial production line may be classified as either "satisfactory" or "defective." With reference to a disease under investigation, a person drawn from a population may be classified as either "has the disease" or "does not have the disease." In general, for many investigations concerning a certain given characteristic, the point of interest is whether an observation reveals the presence of the characteristic or its absence. Since there are only two possible outcomes, they are always of the form "category A" and "category not-A." To simplify terminology, one often uses in a general sense the terms "success" (S) and "failure" (F)—which frequently apply literally to the games of chance around which much of the early development of probability theory took place. Even more convenient are the labels 1 and 0, which have the added advantage that the sum of the observations is the total number of "successes" observed in n trials. The probability structure is completely defined by assigning to the two outcomes S and F (or 1 and 0) two nonnegative numbers totaling unity; it has become customary to designate these probability numbers by p and q, subject to $p + q = 1$. A trial having such a probability structure is called a *Bernoulli* trial, named for the great seventeenth-century mathematician Jacob Bernoulli, who made extensive investigations in this area.

When an experiment consists of n *independent repeated Bernoulli trials*, every possible result of the experiment is a sequence of n characters each of which is S or F (or 1, 0), such as {SFSSSF} in the case $n = 6$. We can talk of the n-*tuple* $(y_1, y_2, \ldots, y_n)$ with each y_i being 1 or 0. Thus {SFSSSF} is the 6-tuple $(1, 0, 1, 1, 1, 0)$. A matter of primary importance in such trials is the number of "successes" in the n trials. In our example with $n = 6$, the number of successes is 4.

It is obvious that the *possible* number of successes in n trials is 0 or 1 or 2 or 3 or $\cdots$ or n. In $n = 6$ trials, we can have 0, 1, 2, 3, 4, 5, or 6 successes. Thus the number of successes is a *variable*, and we say that it is a *random* variable because the specific value depends on chance.

We can see how the pattern of probability for this kind of random variable is developed by considering the simple case of tossing a coin three times, as in Example 4.2.2. If we let $1 =$ head and $0 =$ tail, then all possible fundamental outcomes of the experiment are given by the triples:

$$(1, 1, 1), (1, 1, 0), (1, 0, 1), (1, 0, 0), (0, 1, 1), (0, 1, 0), (0, 0, 1), (0, 0, 0).$$

Each outcome is the result of three independent repeated Bernoulli trials in which $P(\text{success}) = (1/2)$. Thus, each fundamental outcome has probability $(1/2)(1/2)(1/2) = 1/8$. An *event* having to do with the *number of heads* in the three tosses is made up of the appropriate fundamental outcomes. For example, "two heads in three tosses" is the event that occurs if the fundamental outcome is either $(1, 1, 0)$, $(1, 0, 1)$, or $(0, 1, 1)$. We can say that the event "two heads in three tosses" is the *set* of fundamental outcomes composed of the outcomes $(1, 1, 0)$, $(1, 0, 1)$, and $(0, 1, 1)$. It is customary to use braces to indicate such sets, and we can write

$$\text{two heads in three tosses} = \{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}.$$

The probability of such an event is the total probability that goes with the collection of fundamental outcomes. Thus we have

$$P(\text{two heads in three tosses}) = P(\{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}.$$

Considering all possible numbers of heads, we set down the various likelihoods:

$$P(\text{no head in three tosses}) = P(\{(0, 0, 0)\}) = \frac{1}{8},$$

$$P(\text{one head in three tosses}) = P(\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}) = \frac{3}{8},$$

$$P(\text{two heads in three tosses}) = P(\{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}) = \frac{3}{8},$$

$$P(\text{three heads in three tosses}) = P(\{(1, 1, 1)\}) = \frac{1}{8}.$$

We can summarize the entire pattern by using Y to stand for "the number of heads in three tosses," y to represent various specific values of Y, and then write down a table:

y	$P(Y=y)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

$Y = $ number of heads in 3 tosses of a coin

As demanded by logic, the sum of the probabilities in the table is 1, since $Y=0$, $Y=1$, $Y=2$, and $Y=3$ exhaust the totality of all possible outcomes for Y.

Rather obvious changes in the argument will take care of an *unbalanced* coin for which we know the probability of head. As example, suppose $P(\text{head}) = (1/4)$ in any single toss, and we again take three independent tosses. Then the pattern develops as follows.

Outcomes: $(1, 1, 1), (1, 1, 0), (1, 0, 1), (1, 0, 0),$

$(0, 1, 1), (0, 1, 0), (0, 0, 1), (0, 0, 0)$

Probabilities: $\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}, \frac{1}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}, \frac{1}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}, \frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4},$

$\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}, \frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}, \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}, \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}$

| y	Event $Y = y$	$P(Y=y)$
0	$\{(0, 0, 0)\}$	$\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}=1\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^3=\frac{27}{64}$
1	$\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$	$\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{3}{4}+\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}+\frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}=3\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^2=\frac{27}{64}$
2	$\{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}$	$\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{3}{4}+\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{1}{4}+\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}=3\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^1=\frac{9}{64}$
3	$\{(1, 1, 1)\}$	$\frac{1}{4}\cdot\frac{1}{4}\cdot\frac{1}{4}=1\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^0=\frac{1}{64}$

It is a small step to generalize this to cover the case of any given probability of heads: wherever we have had 1/4, use the general value p; wherever 3/4 appears, use q (which is $1-p$). A bit more effort is needed to generalize from three tosses to any desired number, say n tosses. There is indeed a general formula, and it is the basis of calculating probabilities for this large class of situations. For our purposes, however, some representative cases shown by tables will satisfy our needs.

The important notion about Bernoulli trials is the idea of how that mathematical model fits a wide variety of chance situations. It is applicable in every case where the center of interest can be interpreted as

$Y =$ the number of "successes" in n independent trials
of an experiment wherein the probability
of success in a single trial is p.

The random variable identified in this manner is called a *binomial* random variable, and the pattern of its probability structure is called the *binomial distribution*.

For example, the patterns of the probability structure for the cases of tossing a coin three times worked out above are shown graphically as follows:
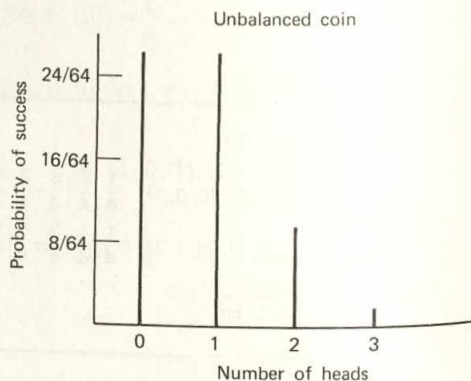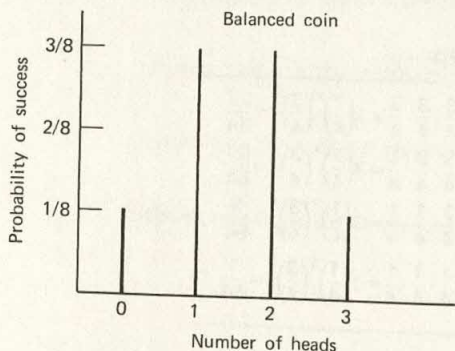
The parameters of the binomial distribution are n and p, a specific numerical distribution being produced for each specific set of values assigned to n, p. The above case of the balanced coin has $n = 3$, $p = (1/2)$; the case of the unbalanced coin is given by $n = 3$, $p = (1/4)$. Table A-2 in the Appendix gives some representative cases. Because practical problems involve a bracket of y values more often than single individual values, published tables usually give their information in the form of cumulative sums. Table A-2 gives such cumulative sums for the number of successes up to and including the tabled y value. For example, the entries for $n = 5$ and $p = .40$ read as follows:

y	p = .40
0	.0778
1	.3370
2	.6826
3	.9130
4	.9898
5	1.0000

The heading of Table A-2 includes the statement "For designated values of n and p, the tabled entry gives $P(Y \leqslant y)$." Thus the above entries stand for the following:

$$P(Y = 0) = \qquad .0778$$
$$P(Y = 0 \text{ or } 1) = \qquad .3370$$
$$P(Y = 0 \text{ or } 1 \text{ or } 2) = \qquad .6826$$
$$P(Y = 0 \text{ or } 1 \text{ or } 2 \text{ or } 3) = \qquad .9130$$
$$P(Y = 0 \text{ or } 1 \text{ or } 2 \text{ or } 3 \text{ or } 4) = \qquad .9898$$
$$P(Y = 0 \text{ or } 1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ or } 5) = 1.0000$$

Example 4.5.1

In a certain very large population it is hypothesized that 30 percent of the individuals need dental treatment. If 10 persons are drawn at random from the population and examined, what is the probability that the number of persons found to need dental treatment *will not exceed* 3, assuming the hypothesis to be correct? Here the large size of the population makes it reasonable that we consider the 10 persons to be 10 independent trials of a situation wherein the probability of "dental treatment needed" is 0.30 at each trial. Thus if $Y =$ the number of persons needing dental treatment, Y can be considered a *binomial* random variable with $n = 10$ and $p = .30$. From Table A-2 we answer the stated question as follows.

$$P(Y \leqslant 3) = .6496.$$

The probability of finding *exactly 3* in need of dental treatment can be calculated as

$$P(Y=3) = P(Y \leqslant 3) - P(Y \leqslant 2) = .6496 - .3828 = .2668.$$

The probability of finding *at least 6 (6 or more, as many as 6)* in need of dental treatment comes from the table as

$$P(Y \geqslant 6) = P(\text{opposite of } Y \leqslant 5) = 1 - P(Y \leqslant 5) = 1 - .9527 = .0473.$$

You may have noticed that the entries in Table A-2 do not include any values of p greater than .50. This is generally the case with published tables of binomial distributions. It is understood that any binomial situation involving a value of p greater than 1/2 can be rephrased in terms of counting the number of "failures," for which the probability at each trial is $q = 1 - p$, giving a binomial distribution with probability parameter less than 1/2.

Example 4.5.2

Suppose a machine has a probability 0.9 of operating successfully during a day's shift, and its day-to-day operations are independent. What is the probability that it will operate successfully at least 13 days out of the next 15? We can argue as follows.

$$Y = \text{the number of successful days;}$$
$$Y \text{ is binomial, } n = 15, \ p = 0.9.$$
$$U = \text{the number of unsuccessful days;}$$
$$U \text{ is binomial, } n = 15, \ p = 0.1.$$
$$P(Y \geqslant 13) = P(U \leqslant 2) = 0.8159.$$

The probability that the machine would operate successfully *no more than 10 days* can be calculated as follows.

$$P(Y \leqslant 10) = P(U \geqslant 5) = P(\text{opposite of } U \leqslant 4)$$
$$= 1 - P(U \leqslant 4) = 1 - .9873 = .0127.$$

EXERCISES

4.5.1 What is the probability of drawing five spades from a deck of bridge cards, if: (a) each draw is replaced and the deck reshuffled before the next draw, (b) each draw is kept out of the deck?

4.5.2 What are the probabilities of drawing five cards *of the same suit* under conditions (a) and (b) of Exercise 4.5.1?

4.5.3 In Section 4.4 we consider the Mecca College data distributed jointly as to sex and political party preference. We found $P(MD) \neq P(M) \cdot P(D)$, and thus concluded that sex and political party preference are not in general independent. Do any of the pairs of categories show independence? (That is, check out the other seven combinations MR, MO, MN, FD, FR, FO, FN.)

4.5.4 In the Mecca College data, are *no opinion on legalizing marijuana* and *no political-party preference* independent?

4.5.5 A seed company claims that 70 percent of a certain kind of seed germinate. You plant 20 of the seeds and find that 12 sprout. What is the probability that no more than this number will sprout if the company's claim is correct?

4.5.6 Suppose the probability is 1/5 that a certain type of missile will arrive and function properly in an assigned target zone. If it does arrive and function properly in the target zone, the target is destroyed.
   a. If five missiles are launched against one target, what is the probability that exactly one missile will arive and function properly in the target zone?
   b. If five missiles are launched against one target, what is the probability that the target will be destroyed?
   c. If 20 missiles are launched against one target, what is the probability that at least 1/5 of them (i.e., at least four) will arrive and function properly in the target zone?

4.5.7 Last year a certain school-bond referendum was defeated, the proportion of "yes" votes being 40 percent. In a recent attitude survey, 15 voters were questioned. Three said they are in favor of a school-bond issue now, and the other 12 registered disapproval. What is the probability of a response as small as this if the general attitude is the same as last year and the 15 voters questioned are a random sample of the corresponding population? If the probability is small, what questions does it raise about interpreting the result of the survey?

4.5.8 In a certain multiple-choice test, each question offers four options for answer; one and only one of the four options is the correct answer. The test has 20 questions. If a person taking the test chooses each answer by pure guessing, what is the probability that he will pass the test (a) if the lowest passing performance is 10 correct answers, (b) if the test is scored by subtracting the number of wrong answers from the number of correct answers, and then taking 10 as the lowest passing score?

4.5.9 The tennis courts in Park A are playable 80 percent of the days in a year. If 10 contest days are chosen, what is the probability that games can be played on the courts at least 8 of the days, assuming the binomial distribution to be applicable? What can you see as an argument against the applicability of the binomial distribution?

4.5.10 Suppose you complain about your experience with the seeds in Exercise 4.5.5, and are told that the probability you computed is large enough to allow the argument that chance has been at work while 70 percent is indeed the germination rate. You ask how small the probability has to be for ruling out that argument, and the reply is "well, under 5 percent." What then is the number of germinating seeds (out of 20) that will send you back to the complaint department?

4.6 PATTERNS OF CHANCE

In the preceding section we considered a probability structure covering a wide variety of chance situations by treating the essential feature as a random variable having the binomial distribution. In general, the use of probability theory goes this way—translate the chance situation into a probability setup having a definable random variable, and then find and apply the probability distribution of that random variable.

Sometimes the case is a very special one. Rolling dice gives an example. Here the interest generally centers on the "point" that is rolled—that is, the sum of the numbers of dots showing on the top faces when the dice come to rest. We have seen earlier how to calculate probabilities for the different "points." The entire structure can be computed and set down in a distribution table, where $Y =$ the point rolled in a fair roll of a pair of fair dice.

y	2	3	4	5	6	7	8	9	10	11	12
$P(Y = y)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Of more general interest are the discovery and use of probability distributions covering broad classes of situations. This is the customary drive behind all mathematics. Whenever we apply any kind of mathematics to the real world, we must use a mathematical pattern or structure—*mathematical model*, it is called in general—which fits the practical situation at hand. Certain models fit broad classes of situations. Thus $A = \pi r^2$ gives the area of any circle, $y = mx + b$ represents any straight line in a plane, $a^2 + b^2 = c^2$ relates the length c of the hypotenuse of any right triangle to the lengths a and b of the other two sides, there is a mathematical formula that fits all compound interest calculations, and there is a mathematical structure that fits the path of any spaceship. So it is with chance situations: there are certain random variables that fit broad classes of cases in the real world. One of these is the *binomial* random variable; we shall introduce a few others that have wide application.

We have seen how the probability distribution of a random variable can be given by a table of probability values. Sometimes a random variable can have its probability values given by a formula—you plug in the value of y and the formula gives you an answer which is $P(Y = y)$. Cases of these kinds can be shown graphically by a *histogram*. In a histogram, a horizontal scale (axis) shows the possible values y, and vertical bars give the probabilities $P(Y = y)$ by height. Each bar is one unit wide, and is centered on the y-value to which it refers. The dice-roll random variable and the binomial random variable of Example 4.5.1 could have their distributions shown as follows.
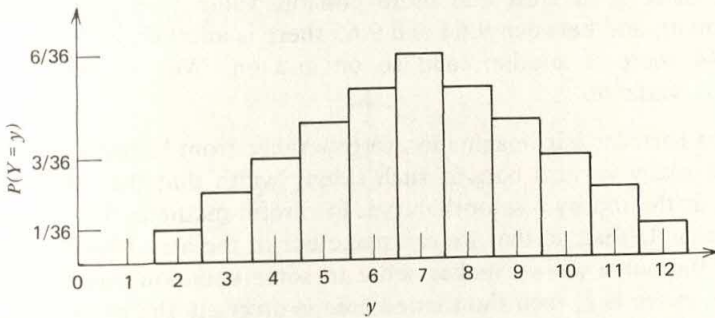
**Probability distribution of point Y rolled with pair of dice.**
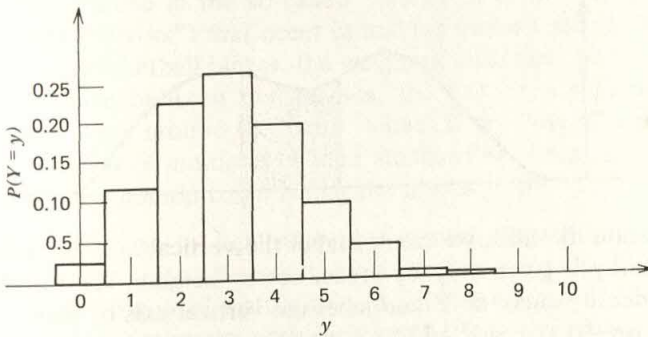
Probability distribution of number Y who need dental treatment when Y is binomial with $n = 10$, $p = 0.3$.

Note that in such histograms we have a correspondence between area and probability. Each bar has a width of 1 and a height of $P(Y = y)$. Hence the bar has area $1 \times P(Y = y) = P(Y = y)$. The total area in the histogram is then the sum of all the probabilities, namely 1.

This is a useful correspondence when we move on to random variables that have too many possible values to allow a histogram representation of their probability distributions. Take, for example, a case of $Y =$ the weight (in pounds) of a child at birth. The possible values of weight are infinite in number, of an especially potent kind of infinity. For between any two values, say 9.6 and 9.7, there is at least one more possible value taking account of finer measurement, and between 9.64 and 9.65 there is another, and between 9.643 and 9.644 there is another, and so on and on. We say that weight is a *continuous* variable.

It is not hard for our imagination to generalize from histograms to diagrams having so many vertical bars of such skinny width that the graph is an area bounded at the top by a smooth curve. In careful mathematical detail the end result is exactly that, so that we can make use of the area idea in very general cases. If the curve shown makes sense in some situation, and the total area under the curve is 1, then the shaded area is precisely the probability that the random variable Y will take on some value between a and b.
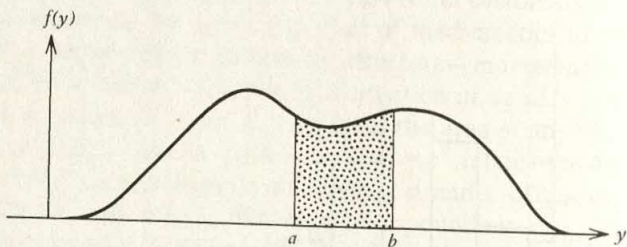
In a situation like this, we cannot label the vertical axis as probability since now probability is given *only* by areas, not by heights. We call the curve the *probability density* curve of Y and label the vertical axis by some notation like the one shown—$f(y)$, read "eff of y", standing for the mathematical function of y which the curve graphs. Such a function is called a *probability density function* and either its equation or its graph gives the story; its *use* is given by taking areas. In most practical situations, published tables allow us to get our desired areas without drawing pictures and measuring.

4.7 THE STANDARD NORMAL DISTRIBUTION

The probability distribution with the widest use among continuous random variables is a special one given the name *standard normal*. It is sometimes referred to also as the *Gaussian* distribution, named for the German mathematician Karl Friedrich Gauss (1777–1855), who had much to do with its derivation and early use in the study of measurement errors.

The graph of the *standard normal* probability density function is the so-called "bell-shaped curve":
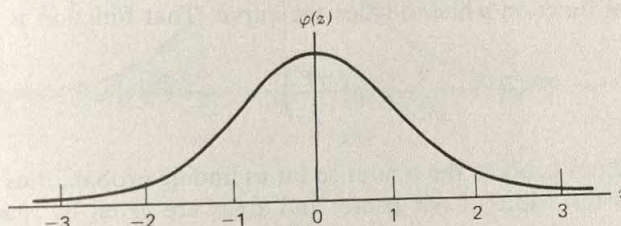
The standard normal distribution.

This distribution arose in the so-called "theory of errors" by study of the random variations ("errors") that occur in making measurements. If we measure the height of a basketball player, the weight of a sleeping bag, the speed of a bullet, the distance between two planets, the I.Q. of a student, repeated measurements will vary around the "true" value. If the "overs" and "unders" are expressed in terms of multiples of their standard deviation, the probability pattern in the most common cases is like the above graph.

The distribution came to light also in the early searches made to find approximations to the probabilities in the binomial distribution. Even the brief discussion about the binomial distribution in Section 4.5 must have suggested to the reader that calculating the probabilities must be a real chore whenever n is large. That is indeed a fact, and so it was a major discovery (De Moivre, 1733) that the binomial distribution has a closer and closer relationship to the above graph as n gets larger and larger. We shall see how this works in Section 4.9.

In the years since its formulation, the standard normal distribution has been found useful for describing the probability patterns of a wide variety of random variables of the continuous type. It has been the hardest worked, and sometimes most badly overworked, of all known probability distributions. The nomenclature *normal* has nothing to do with being the opposite of "abnormal." The term arose apparently through the notion that when a random variable is expressed as a directed distance of so many standard deviations away from the mean, it has been "standardized" or "normalized."

This distribution appears in so many various applications that we set aside notation for its own use: Z for the standard normal random variable, z for the numerical variable indicating the values that Z can assume, and $\varphi(z)$ for the mathematical function which defines the curve. That function is specifically

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right), \qquad -\infty < z < \infty,$$

but this need not concern the reader so far as finding probabilities is concerned, since *area* is the name of our game, and areas are given by readily available tables.

Note that the standard normal curve is defined to go along off to the left as far as "minus infinity" and off to the right as far as "plus infinity." Obviously no one will ever get to see either end, and a logical purist can raise the objection that the standard normal distribution cannot match anything in real life since nothing measurable involves the infinitely negative or the infinitely positive. But you can see from the graph that the area under the curve beyond $z = -3$ on the left or beyond $z = +3$ on the right is very small. In the next paragraph we shall see how small. Thus as in so much of mathematics, the standard normal probability distribution can *approximate* reality since the probability shown for unrealistic values of z is so small as to be "zero for all practical purposes."

Table A-3 is an abridged table of the standard normal distribution. It is arranged in such a way as to be most convenient for general use. Tables giving more detailed entries are widely available in textbooks, mathematics handbooks, and separate volumes.

Shaded area $= P(Z \leqslant z^*)$

Area = .1587

$P(Z \leqslant -1.0) = .1587$

Area = .7258

$P(Z \leqslant 0.6) = .7258$
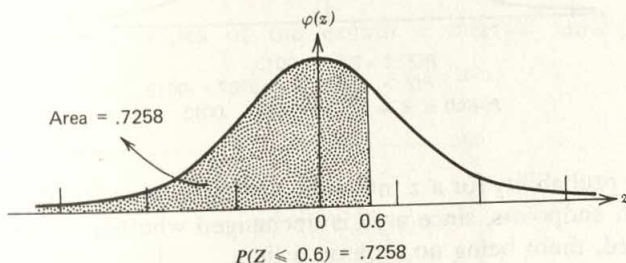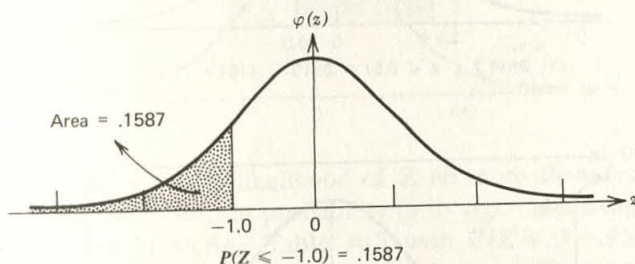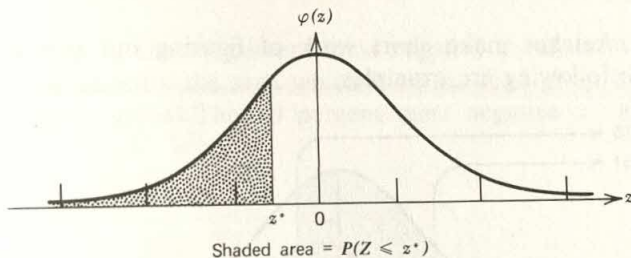
When we refer to Table A-3 with a value of z, the corresponding reading is the *area under the curve to the left of that z value*. The column is labeled $P(Z \leqslant z)$ since that is what the indicated area represents. By obvious maneuvers we can apply such pieces of information to produce the probability of any desired interval of z values. For example, the last two diagrams give immediately

$$P(Z > -1.0) = 1 - P(Z \leqslant -1.0) = 1 - .1587 = .8413,$$
$$P(Z > 0.6) = 1 - P(Z \leqslant 0.6) = 1 - .7258 = .2742.$$

Abbreviated sketches make short work of figuring out probabilities for z intervals. The following are examples.

$P(-1.2 < Z \leqslant 0.5) = .6915 - .1151 = .5764$

$P(Z < -3.0) = .0013,$
$P(Z > +3.0) = 1 - .9987 = .0013,$
$P(-3.0 \leqslant Z \leqslant +3.0) = .9987 - .0013 = .9974.$
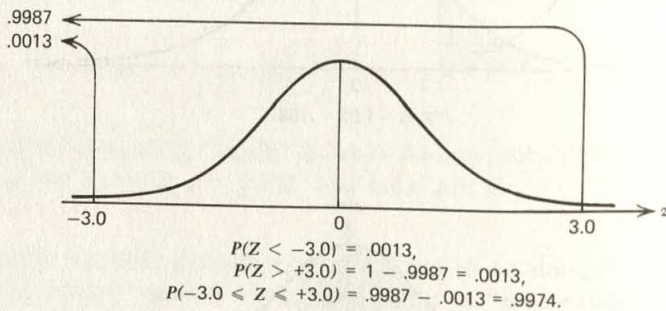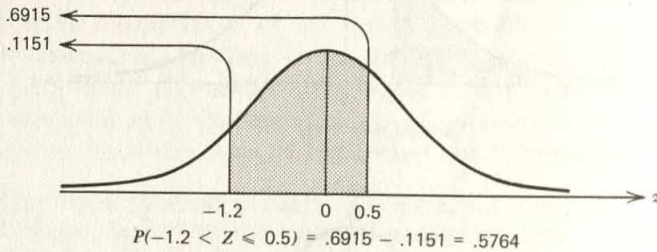
Note that the probability for a z interval is the same whether or not we include either or both endpoints, since area is unchanged whether or not a boundary line is included, there being no area on a line.

By using Table A-3 in reverse fashion, we can answer questions about z values which bound any desired likelihood region of Z. The following are examples.

Given: area .10

-1.282 0
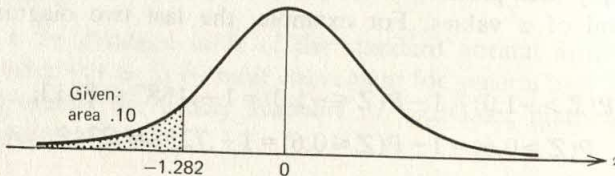
What is the limit of the 10 percent most negative z values? We graph the situation at hand, identify the area which is of the kind given by Table A-3, and read off the z value. The 10 percent most negative z values are $z \leqslant -1.282$.
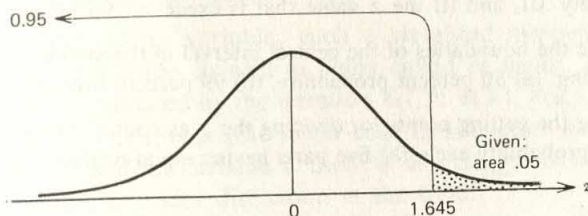
Beyond what z value is the likelihood of Z no more than 5 percent? The questioned z value has 5 percent probability to its right, and hence 95 percent probability to its left. Entering Table A-3 with $P(Z \leqslant z) = .9500$, we read $z = 1.645$. The diagram shows that any z value from 1.645 on out to the right has 5 percent or less probability beyond it.

What are the boundaries of the *central* z interval having 95 percent probability?

$$P(-1.960 < Z < 1.960) = .95$$
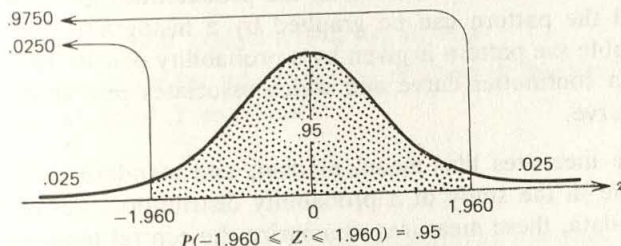
EXERCISES

4.7.1 Find the probability of each of the following statements of behavior of the standard normal random variable Z: (a) $Z < -0.8$, (b) $Z < 2.0$, (c) $Z > 1.0$, (d) $Z > -0.2$, (e) $Z \leqslant 1.2$, (f) $Z \leqslant -1.2$, (g) $Z \geqslant 1.2$, and (h) $Z \geqslant -1.2$.

4.7.2 Find the following probabilities concerning the standard normal random variable Z: (a) $P(-0.6 < Z < 1.6)$, (b) $P(1.6 < Z < 2.6)$, (c) $P(-1.9 < Z < -0.9)$, (d) $P(-1.5 \leqslant Z \leqslant +1.5)$, (e) $P(Z \leqslant -1.5$ or $Z \geqslant +1.5)$, and (f) $P(Z \leqslant -1.5$ or $Z \geqslant +2.0)$.

4.7.3 Identify the value of z in each of the following statements about the standard normal distribution: (a) $P(Z \leqslant z) = .10$, (b) $P(Z < z) = .70$, (c) $P(Z > z) = .70$, (d) $P(Z > z) = .10$, (e) $P(Z \leqslant z) = .50$, (f) $P(Z \geqslant z) = .50$, (g) $P(Z < z) = .25$, (h) $P(Z > z) = .80$, and (i) $P(-z < Z < z) = .90$.

4.7.4 Identify the values of z satisfying the following conditions: (a) the upper limit of the values of Z in the lowest 15 percent of the distribution, (b) the lower z limit of the upper 25 percent of the distribution, (c) the z value that has 60 percent probability to its left on the z axis, (d) the z value that has 92 percent probability to its right on the z axis, (e) the z value that is exceeded with probability .01, and (f) the z value that is exceeded 95 percent of the time.

4.7.5 What are the boundaries of the central interval in the standard normal distribution having: (a) 80 percent probability, (b) 99 percent probability?

4.7.6 What are the cutting points for dividing the z axis into: (a) four parts having 25 percent probability each, (b) five parts having equal probability?

4.8 DESCRIPTIVE MEASURES IN PROBABILITY DISTRIBUTIONS

In Chapters 2 and 3 we considered various ways of showing the pattern exhibited by a set of data, and the most common summarizing statistics to indicate central tendency and dispersion. For central tendency we dealt with the mean $\bar{y}$ or the median or the mode of the data, and for dispersion we made use of the range or the variance s^2 or the standard deviation s.

Similar considerations are useful with regard to the probability distribution of a random variable. The pattern has been discussed in preceding sections. For a discrete random variable we have all the probabilities given by a table or a formula, and the pattern can be graphed by a histogram. For a continuous random variable the pattern is given by a probability density function, which is graphed by a continuous curve and which associates probabilities with areas under that curve.

Descriptive measures like mean, median, and standard deviation play an important role in the study of a probability distribution. As in the case of a collection of data, these measures summarize the central tendency and dispersion of the random variable. In most situations of practical interest they enable us to identify the arbitrary constants (parameters) in a distribution of known form and thus to identify the probability pattern completely.

Recall the definition of *sample mean*: $\bar{y} = \sum y/n$. This can be written as

$$\bar{y} = \frac{\sum y}{n} = y_1\left(\frac{1}{n}\right) + y_2\left(\frac{1}{n}\right) + \cdots + y_n\left(\frac{1}{n}\right).$$

In this form we see $\bar{y}$ as a *weighted average* of the ys: each y-value is given the weight $(1/n)$, we multiply ("weight") each y-value by that amount $(1/n)$, and add the results. This total is then divided by the sum of the weights, but here the sum of the weights is just 1:

$$\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n} = n\left(\frac{1}{n}\right) = 1.$$

The mean of a *random variable* Y (we call it also the mean of the *probability distribution* of Y) is defined in a similar manner. But now the y values at hand are *all the possible values* that Y can assume, and the weights are the probabilities associated with those various values. Again the sum of the weights is 1, since the total probability mass ("weight") in any chance situation is 1.

For any kind of random variable, such a weighted average is called its *expectation* or *expected value*, as well as being called its *mean*. The process of taking expectation is indicated by the notation $\mathscr{E}(\square)$: $\mathscr{E}(Y)$, $\mathscr{E}(Z)$, $\mathscr{E}(Y^2)$, and so on. The shorthand notation is μ (the lower-case Greek letter *mu*), subscripted to show the related random variable if there is any danger of confusion as to what random variable is under discussion at the time.

> If Y is a *discrete* random variable, its probability distribution is given by the probability function $f(y) = P(Y = y)$, and the mean of Y can be readily defined by complete formula:
>
> $$\left.\begin{array}{l}\text{Mean} \\ \text{Mean value} \\ \text{Expected value} \\ \text{Expectation}\end{array}\right\} \text{ of } Y = \mu = \mathscr{E}(Y) = \sum_{\text{all } y} y \cdot P(Y = y) = \sum_{\text{all } y} y \cdot f(y). \quad (4.8.1)$$

Example 4.8.1. Y = "*point*" *rolled with a pair of dice.*

The probability function for this Y has been given in Section 4.6. We now can calculate the mean of Y as follows.

y	$f(y) = P(Y = y)$	$y \cdot f(y)$
2	1/36	2/36
3	2/36	6/36
4	3/36	12/36
5	4/36	20/36
6	5/36	30/36
7	6/36	42/36
8	5/36	40/36
9	4/36	36/36
10	3/36	30/36
11	2/36	22/36
12	1/36	12/36
Any other value	0	0
		$\mu = \mathscr{E}(Y) = (252/36) = 7$

This is shown on the graph of the probability distribution in the diagram below, which shows that the center of gravity of this binominal distribution is $\mathscr{E}(Y) = \mu = 7$.

Probability distribution of point Y rolled with pair of dice

$\mu = \mathscr{E}(Y) = 7$
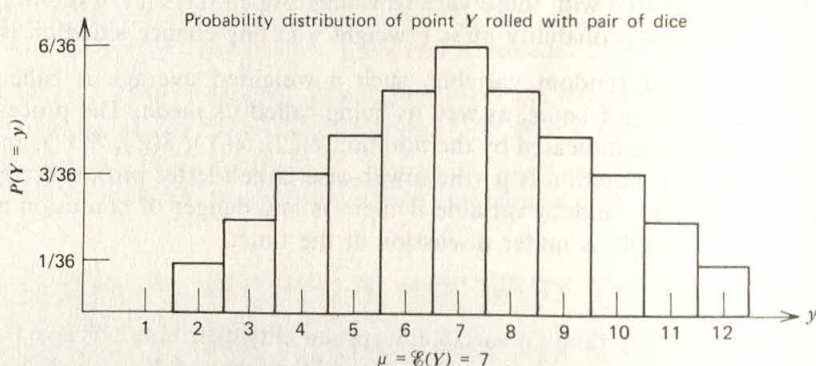
When Y is a *continuous* random variable, its probability distribution cannot be given by a probability function of the kind that takes care of a discrete random variable. We have instead a probability *density* function, and probabilities are given by areas under the graphed curve. In such a case we cannot find a weighted average of the y values by a simple summing process like the one above. Methods of the *calculus* are required, performing a particularly tricky kind of summation yielding a result having the same concept as above—a probability-weighted average of all possible y values. We shall set down the formula for this, but *only for the purpose of completing our catalogue*. The reader will *never in this book* be asked to calculate this formula—unless he is unwilling to take our word for the answer which we shall give wherever needed!

If Y is a *continuous* random variable, having probability density function $f(y)$, then the mean μ of Y is

$$\mu = \mathscr{E}(Y) = \int_{-\infty}^{\infty} y \cdot f(y) \, dy. \tag{4.8.2}$$

It is a point of interest in the history of mathematics that the symbol ∫ above (called the *integral sign*) was devised by stretching out the capital letter S, standing for *sum*.

In Chapter 3 we took for our descriptive measure of dispersion in a collection of data the *sample standard deviation s*, found by taking the positive square root of the *sample variance s²*, using one or another computing form of the formula

$$s^2 = \frac{\sum (y - \bar{y})^2}{n-1}, \qquad s = +\sqrt{s^2}.$$

For the probability distribution of a random variable Y we use the same concept of averaging squared deviations of y values from their mean. Here the mean is the mean μ of Y, instead of the mean $\bar{y}$ of the sample data, and the averaging is according to probability "weights." We use the notation σ^2 for variance and σ for standard deviation. Sometimes $V(Y)$ or $\text{var}(Y)$ is used to denote variance of Y.

If Y is a random variable having probability function $f(y)$ or probability density function $f(y)$, the *variance of Y* (called also the *variance of the distribution of Y*) is denoted by σ^2 or $V(Y)$ or $\text{var}(Y)$, and is defined as

$$\sigma^2 = V(Y) = \mathscr{E}([Y - \mu]^2) = \begin{cases} \sum_{\text{all } y} (y - \mu)^2 f(y) & \text{if Y is discrete,} \\ \int_{-\infty}^{\infty} (y - \mu)^2 f(y) dy & \text{if Y is continuous.} \end{cases} \qquad (4.8.3)$$

Again the reader should look at the formula for the continuous case as only a specialized kind of summation which he will not be asked to calculate.

The *standard deviation of Y* (called also the *standard deviation of the distribution of Y*) is σ, the positive square root of the variance σ^2:

$$\sigma = +\sqrt{\sigma^2}. \qquad (4.8.4)$$

Notice that our notation is consistent with the following useful distinction between *population* and *sample* parameters:

	Population	Sample
Mean	μ	$\bar{y}$
Variance	σ^2	s^2
Standard deviation	σ	s

Example 4.8.2. Y = "point" rolled with a pair of dice.

In Example 4.8.1 we found μ to be 7 for this Y. The variance and standard deviation of Y then come out as follows.

y	f(y)	y−μ	(y−μ)²	(y−μ)² · f(y)
2	1/36	−5	25	25/36
3	2/36	−4	16	32/36
4	3/36	−3	9	27/36
5	4/36	−2	4	16/36
6	5/36	−1	1	5/36
7	6/36	0	0	0
8	5/36	1	1	5/36
9	4/36	2	4	16/36
10	3/36	3	9	27/36
11	2/36	4	16	32/36
12	1/36	5	25	25/36

$$\sigma^2 = 210/36 = 35/6 = 5.83$$

variance of $Y = \sigma^2 = 5.83$; and
standard deviation of $Y = \sigma = \sqrt{5.83} = 2.41$.

Example 4.8.3. The binomial distribution.

In the general case of this distribution, where Y = the number of "successes" in n independent trials of an experiment wherein the probability of success in a single trial is p, the appropriate mathematics (with which we don't want to bother you) will show:

$$\text{Binomial distribution} \begin{cases} \text{Mean:} & \mu = np \\ \text{Standard deviation:} & \sigma = \sqrt{np(1-p)} \end{cases} \qquad (4.8.5)$$

We can see an example of the consistency of this with the definitions (4.8.1) and (4.8.3) by using both procedures on the easy case of Y = number of heads in three tosses of a fair coin. Early in Section 4.5 this random variable was

considered, and its probability distribution set down in tabular form. Let us take that now and apply (4.8.1) and (4.8.3):

y	$f(y)$	$y \cdot f(y)$	$y - \mu$	$(y - \mu)^2$	$(y - \mu)^2 \cdot f(y)$
0	$\frac{1}{8}$	0	$-\frac{3}{2}$	$\frac{9}{4}$	$\frac{9}{32}$
1	$\frac{3}{8}$	$\frac{3}{8}$	$-\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{32}$
2	$\frac{3}{8}$	$\frac{6}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{32}$
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{2}$	$\frac{9}{4}$	$\frac{9}{32}$
		$\mu = \frac{12}{8} = \frac{3}{2}$			$\sigma^2 = \frac{24}{32} = \frac{3}{4}$

By use of (4.8.5) we have the identical results:

$$\mu = np = 3(1/2) = 3/2,$$
$$\sigma^2 = np(1-p) = 3(1/2)(1/2) = 3/4.$$

In the situation of Example 4.5.1, where we drew a sample of 10 persons from a population in which 30 percent need dental treatment, Y is *binomial* with $n = 10$ and $p = 0.30$, so that

$$\mu = \text{mean of } Y = np = (10)(.30) = 3.0,$$

$$\sigma = \text{standard deviation of } Y = \sqrt{np(1-p)} = \sqrt{10(.3)(.7)} = \sqrt{2.1} = 1.4.$$

The value np readily matches our intuition on what the "expected" number of successes should be. When we make 10 trials, each with success probability 30 percent, we surely "expect" 30 percent of the 10 trials to be successes *on the average*. This on-the-average expected number is what the mean tells us.

The standard deviation is not so readily comprehended intuitively, but we do know by common sense that the actual *observed* number of successes will *vary* around the mean. The standard deviation is a measure of that variability.

Example 4.8.4. Binomial with $n = 150$, $p = .30$.

Instead of taking a sample of 10 persons from the population of Example 4.5.1, take at random 150 persons. Then the number Y of sampled persons needing dental treatment is *binomial* with $n = 150$ and $p = .30$. Then for this Y we have

$$\mu = np = 150(.30) = 45.0,$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{150(.3)(.7)} = \sqrt{31.5} = 5.6.$$

Example 4.8.5.   The machine of Example 4.5.2.

The machine has a probability 0.9 of operating successfully during a day's shift, and its day-to-day operations are independent. Considering Y as the number of successful days out of 15, we have:

$$Y \text{ is binomial, } n = 15, \, p = 0.9;$$

$$\mu = np = 15(.9) = 13.5,$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{15(.9)(.1)} = \sqrt{1.35} = 1.2.$$

If we ask about successful days in the next 90-day period, we have:

$$Y \text{ is binomial, } n = 90, \, p = 0.9;$$

$$\mu = np = 90(.9) = 81.0,$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{90(.9)(.1)} = \sqrt{8.10} = 2.8.$$

In this book we shall not be interested in extensive study of finding the mean and standard deviation of various random variables. That is part of a careful study of probability theory. Here we need just to get straight the concept of the mean and standard deviation of a random variable Y *as distinct from* the mean and standard deviation of a collection of data relating to Y.

When Y is a random variable, it is associated with a chance process. That process has a mean μ and a standard deviation σ in accordance with the probability averaging discussed above. Each time we operate the process we get a specific numerical result—a y value given to us by Nature using the chance mechanism that is associated with the probability distribution of Y. We call such a y value an *observation on* Y, or an *observation on the population of* Y. A collection of such observations is called a *sample from the population of* Y. The number of observations in a sample is called the *size* of the sample.

For most cases of practical interest in statistical analysis, such a sample, no matter how large its size, cannot tell us the complete story about Y—there are always more observations which could be made. Thus the mean $\bar{y}$ of the *sample* can never give us precisely μ, the mean of the *population* of Y, nor can the *sample* standard deviation s give us the *population* standard deviation σ. What we shall do with statistical inference is use $\bar{y}$ and s to make educated guesses about μ and σ.

An essential point to keep in mind about a chance process is that the *most* we can *know* about it is the probability distribution involved. We can never know what a dice roll *will* show; all we can know is the *likelihood* of the various possible results. If Y is a *binomial* random variable, the most that can be known is the pattern of likelihood given by the probability function for the binomial distribution, and to completely know *that* requires knowledge of the specific values for n and p. Similarly for any other kind of random variable, the

totality of knowledge concerning it is given by: (a) the form of its probability function or probability density function and (b) the specific numerical values of the arbitrary constants (parameters) in that function.

In the important case of the *general* normal distribution, which is the generalization of the *standard* normal distribution studied in the preceding section, the mean μ and standard deviation σ completely specify the distribution in accordance with the following diagram.
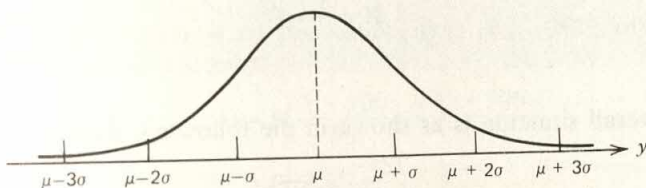
Normal distribution with mean μ and standard deviation σ.

The probability structure is that of the standard normal curve centered at μ instead of at zero, and using σ as the unit of measurement on the horizontal scale. Such a random variable Y is designated *normal, with mean μ and standard deviation σ.*

The shorthand notation $N(\mu, \sigma)$ is often used, so that we can write

$$Y \text{ is } N(\mu, \sigma)$$

to mean "Y has the normal distribution with mean μ and standard deviation σ."

For example, the probability distribution of scholastic aptitude test (SAT) scores in a certain population is $N(500, 50)$ and is shown in the following diagram. Here we see that $\mu = 500$, and $550 = \mu + \sigma$, and so on. In other
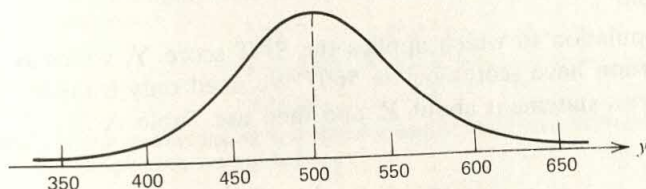
words, 550 is just one standard deviation unit above the mean. Similarly 400 is two standard deviation units below the mean.

It would be very useful to convert all the numbers on the y scale of measurement to the z scale of measurement that we used to discuss the standard normal distribution in Section 4.7. We could use a single table for all normal populations.

The conversion formula gives "the standardized normal random variable Z":

$$Z = \frac{Y - \mu}{\sigma}$$

Thus the overall situation is as shown in the following diagram.
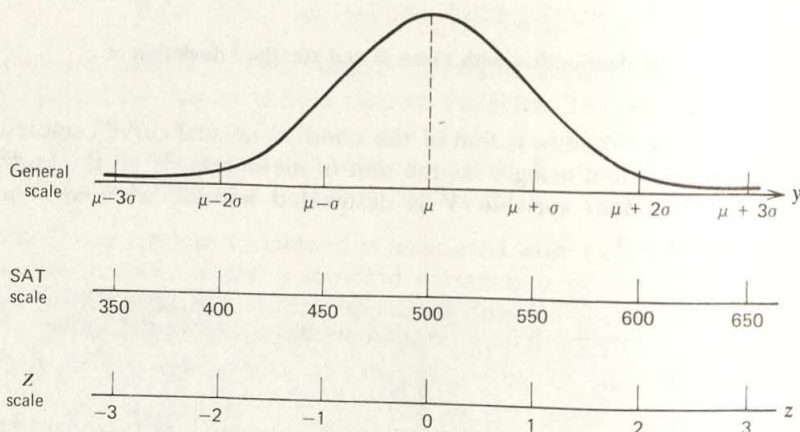
We see that 400 is -2 standard deviation units below μ; 550 is $+1$ unit above μ, and so on. Thus, knowing the mean (μ) and standard deviation (σ) of a normal random variable Y, we can translate back and forth between Y and the standard normal random variable Z by the relationship

$$\frac{Y - \mu}{\sigma} = Z, \quad \text{where } Y \text{ is } N(\mu, \sigma) \text{ and } Z \text{ is } N(0, 1). \tag{4.8.6}$$

Example 4.8.6

In the population to which applies the SAT score Y, which is $N(500, 50)$, what proportion have scores below 560? We need only translate a statement about Y into a statement about Z, and then use Table A-3:

$$P(Y < 560) = P\left(\frac{Y - 500}{50} < \frac{560 - 500}{50}\right)^*$$

$$P\left(Z < \frac{60}{50}\right)$$

$$= P(Z \le 1.2) = .8849.$$

Thus 88.49 percent of the population have SAT scores lower than 560.

Example 4.8.7

In the above population, what proportion have scores 425 or less? What proportion have scores over 600?

$$P(Y \le 425) = P\left(\frac{Y - 500}{50} \le \frac{425 - 500}{50}\right)$$

$$= P\left(Z \le \frac{-75}{50}\right)$$

$$= P(Z \le -1.5) = .0668.$$

Thus 6.68 percent of the population have scores 425 or less.

$$P(Y > 600) = P\left(\frac{Y - 500}{50} > \frac{600 - 500}{50}\right)$$

$$= P\left(Z > \frac{100}{50}\right)$$

$$= P(Z > 2.0)$$

$$= 1 - P(Z \le 2.0)$$

$$= 1 - .9773 = .0227.$$

Thus, 2.27 percent of the population have scores above 600.

---

* The algebra of inequalities is very much like the algebra of equalities, with one additional wrinkle. We operate with the following in mind:
(a) An inequality is unchanged if a common quantity is added to, or subtracted from, each member, for example,

$$5 < 7 \qquad\qquad 5 < 7$$
$$2 + 5 < 2 + 7 \qquad 5 - 8 < 7 - 8$$
$$7 < 9 \qquad\qquad -3 < -1$$

(b) An inequality is unchanged if each member is multiplied or divided by a common *positive* quantity, for example,

$$5 < 7 \qquad\qquad 16 > 12$$
$$2 \times 5 < 2 \times 7 \quad 16 \div 4 > 12 \div 4$$
$$10 < 14 \qquad\qquad 4 > 3$$

(c) An inequality has its sense (direction) reversed if each member is multiplied or divided by a common *negative* quantity, for example,

$$5 < 7 \qquad\qquad 16 > 12$$
$$(-2) \times 5 > (-2) \times 7 \quad 16 \div (-4) < 12 \div (-4)$$
$$-10 > -14 \qquad\qquad -4 < -3$$

Example 4.8.8

In the same population as in the two preceding examples, what is the top score of the lowest 15 percent? We start with Table A-3, enter with area, read the z value and then work out y.

Area .15

−1.036   0
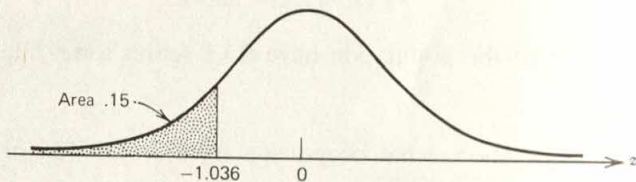
$$.15 = P(Z \leqslant -1.036)$$
$$= P\left(\frac{Y - 500}{50} \leqslant -1.036\right)$$
$$= P(Y - 500 \leqslant 50(-1.036))$$
$$= P(Y \leqslant 500 - 51.8)$$
$$= P(Y \leqslant 448.2)$$

Thus the lowest 15 percent of SAT scores in the population are 448 and below.

Example 4.8.9

In the above population, what are the scores for the top 5 percent?

.9500

Area .05
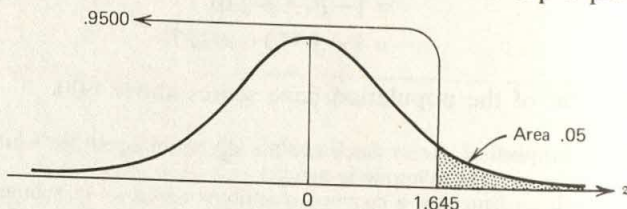
0   1.645

$$.05 = P(Z \geqslant 1.645)$$
$$= P\left(\frac{Y - 500}{50} \geqslant 1.645\right)$$
$$= P(Y - 500 \geqslant 50(1.645))$$
$$= P(Y \geqslant 500 + 82.25)$$
$$= P(Y \geqslant 582.25)$$

Thus the top 5 percent of the population have SAT scores of 582 and over.

In Chapter 3 we considered the usefulness of the *median* as a descriptive measure of central tendency for a collection of data. For such a set of observations the median is the "middle" observation, or the mean of the two "middle" observations, *in order of magnitude;* it is thus a scale value that is exceeded by as many observations as it exceeds. With respect to a *random variable* Y and its probability distribution, we translate this notion into probability terms, just as we translated the notions of mean and standard deviation. We would like to define the *median of* Y as that y value that is *as likely to be* exceeded as not.

In almost every case where Y is a *discrete* random variable, there is trouble in trying to identify a median by the above criterion, because the probability accumulates by jumps. Consider the random variable about the point rolled with a pair of dice. In Section 4.6 we set down the entire probability distribution of that Y. The following table shows what happens when we look for a y value that is as likely to be exceeded as not. For each y value we put in one column the total probability up to and including that y and in another column the probability of larger values of y; then we look for a y-value where the two entries are the same.

y	Probability of y and smaller values	Probability of larger values
1	0	1
2	1/36	35/36
3	3/36	33/36
4	6/36	30/36
5	10/36	26/36
6	15/36	21/36
7	21/36	15/36
8	26/36	10/36
9	30/36	6/36
10	33/36	3/36
11	35/36	1/36
12	1	0

We see that there is *no* y value that is as likely to be exceeded as not. Hence we must look for some kind of conventional rules by which we can select some number as a reasonable measure of "break-even" point. We do not want to enter that jungle in this book, and so we shall restrict ourselves to *continuous* random variables when we consider the median, or other scale values for splitting the probability distribution into different ratios.

When Y is a continuous random variable, its entire probability mass is represented by the area under a continuous curve, and so we can precisely find a y value to cut the area at any proportion we want. It is of interest to define a whole set of cutting points for giving *percentages* of the total area. Such cutting points are called *percentiles*.
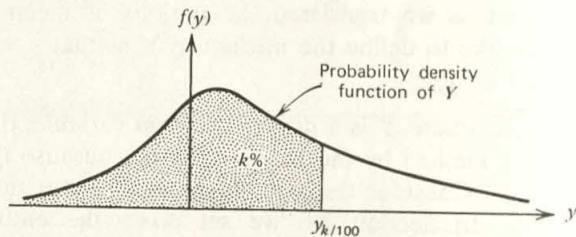
DEFINITION 4.8.1 For any continuous random variable Y, the kth *percentile of* Y (or, of the distribution of Y) is that y-value, denoted $y_{k/100}$, which satisfies the relation $P(Y \leqslant y_{k/100}) = k$ percent.

Thus the *median* of Y is the 50th percentile $y_{.50}$, since

$$P(Y \leqslant y_{.50}) = .50 = 50 \text{ percent}$$

whence

$$P(Y > y_{.50}) = 1 - P(Y \leqslant y_{.50}) = 1 - .50 = .50 = 50 \text{ percent},$$

and, since Y is continuous,

$$P(Y < y_{.50}) = P(Y \leqslant y_{.50}),$$

so that we have

$$P(Y < y_{.50}) = .50 = P(Y > y_{.50}).$$

In some areas of application, special names are given to certain classes of percentiles, according to the following scheme:

$$y_{.25}, y_{.50}, y_{.75}: \quad \text{first, second, third } \textit{quartile};$$
$$y_{.20}, y_{.40}, y_{.60}, y_{.80}: \quad \text{first, second, third, fourth } \textit{quintile};$$
$$y_{.10}, y_{.20}, \ldots, y_{.90}: \quad \text{first, second, } \ldots, \text{ninth } \textit{decile}.$$

In this array we can note that the *median* is the 50th percentile or the second quartile or the fifth decile.

Example 4.8.10   Normal distributions.

Table A-3 gives immediately the percentile rank of the listed z values in the standard normal distribution: we have only to read $P(Z \leqslant z)$ and move the decimal place. Thus for $z = 1.3$ we read $P(Z \leqslant 1.3) = .9032$, and so 1.3 is the 90.32th percentile of Z. Granted, it is not very practical to read this out loud! We can try "ninety point thirtytwoth percentile," but it is perhaps better left unsaid, relying on statements like "somewhat higher than the ninetieth percentile" or "between the ninetieth and ninety-first percentiles."

For a general normal distribution we can find percentile rankings by translating normal Y into standard normal Z, as we did earlier. Take the example of SAT scores which are $N(500, 50)$. What percentile is the score 525?

$$P(Y \leqslant 525) = P\left(\frac{Y - 500}{50} \leqslant \frac{525 - 500}{50}\right)$$

$$= P\left(Z \leqslant \frac{25}{50}\right)$$

$$= P(Z \leqslant 0.5) = .6915.$$

Thus 525 is just above the 69th percentile.

We identify z values or y values for specified percentiles by running the above procedures in reverse. From Table A-3 we see that the 20th percentile of the standard normal distribution is -0.842, since we can enter the table where $P(Z \leqslant z)$ lists .2000 and read out $z = -0.842$. Similarly the 90th percentile of Z is 1.282, the 5th percentile is -1.645, and the median (50th percentile) is 0.
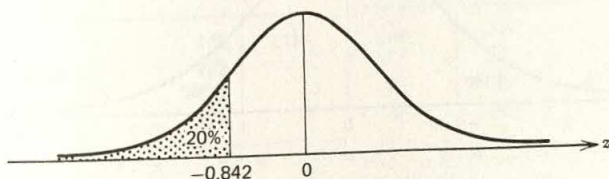
Translation from Z to Y will carry any such argument to the percentiles of any given general normal distribution. Again take the $N(500, 50)$ SAT scores as example. What is the 20th percentile of this distribution?

$$.20 = P(Z \leqslant -0.842)$$

$$= P\left(\frac{Y - 500}{50} \leqslant -0.842\right)$$

$$= P(Y - 500 \leqslant -42.1)$$

$$= P(Y \leqslant 457.9).$$

Thus the 20th percentile of the distribution is approximately the score 458. Similarly

$$.90 = P(Z \leqslant 1.282) = P\left(\frac{Y-500}{50} \leqslant 1.282\right) = P(Y - 500 \leqslant 64.1)$$
$$= P(Y \leqslant 564.1)$$

$$.05 = P(Z \leqslant -1.645) = P\left(\frac{Y-500}{50} \leqslant -1.645\right) = P(Y - 500 \leqslant -82.25)$$
$$= P(Y \leqslant 417.75)$$

show that the 90th percentile is approximately 564 and the 5th percentile approximately 418.

The following diagram is a frequently used description of the normal distribution. We start with the standard normal curve, take cutting points at $z = \pm 1, \pm 2, \pm 3$, and use Table A-3 to get the indicated areas (probabilities). Using the relation (4.8.6):

$$Z = \frac{Y - \mu}{\sigma}, \qquad \text{when } Y \text{ is } N(\mu, \sigma),$$

we can convert the z scale to a y scale for the general normal random variable, since the above equation gives

$$z = \frac{y - \mu}{\sigma}, \qquad \sigma z = y - \mu, \qquad y = \mu + \sigma z.$$
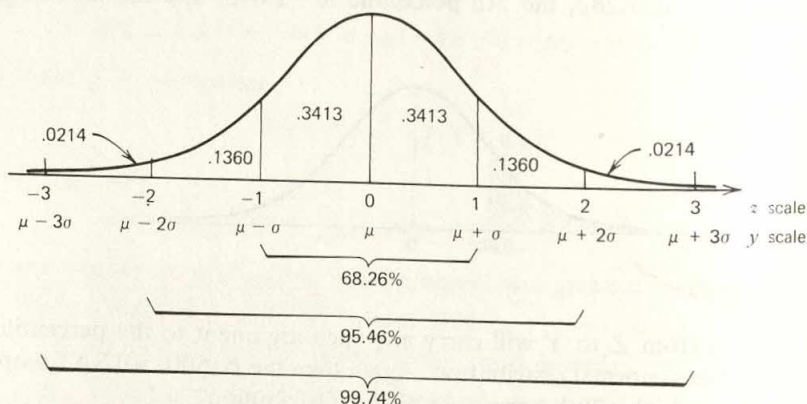
It is because of the facts in this diagram that one frequently hears a statement such as "in a normal population, approximately 68 percent of all of the elements are within one standard deviation of the mean, about 95 percent of them are within two standard deviations of the mean, and all but about one fourth of one percent of them are within three standard deviations of the

mean." Here we should keep in mind that "normal" means normal as defined by the *normal distribution*. And if "population" is a population of physical elements like people or television tubes or soap bars, then the normal distribution can be only an approximation, since any collection of a finite number of elements has a histogram for its distribution graph, as we saw in Chapter 2. To say that such a collection has a "normal distribution" makes sense only if the collection is so large and so distributed that its histogram if plotted on a fine mesh of intervals is virtually indistinguishable from the normal distribution curve.

EXERCISES

4.8.1 If the random variable Y has the following probability distribution, what are the mean (μ) and standard deviation (σ) of Y?

y	$f(y) = P(Y = y)$
1	0.20
2	0.40
6	0.30
12	0.10

4.8.2 Calculate the mean and variance for each of the following random variables:

(a)

y	$f(y) = P(Y = y)$
0	1/8
3	1/2
6	1/8
7	1/4

(b)

y	$f(y) = P(Y = y)$
−2	.25
−1	.30
0	.05
3	.10
5	.30

4.8.3 In Exercise 4.3.2 you calculated the probabilities for all possible numbers of *heads* in four tosses of a balanced coin. You thus constructed the probability distribution for the binomial random variable Y = the number of heads in four tosses (the parameters being $n = 4$ and $p = 1/2$). Use your results from Exercise 4.3.2 to calculate the mean (μ) and variance (σ^2) of Y by the formulas for μ and σ^2 given by (4.8.1) and (4.8.3). Check your results by showing agreement with (4.8.5): the binomial distribution has $\mu = np$, $\sigma^2 = np(1-p)$.

4.8.4 For the random variable Y having the binomial distribution with $n = 10$ and $p = 0.4$, calculate the mean (μ) and variance (σ^2) by the formulas for μ and σ^2 in (4.8.1) and (4.8.3), using the probabilities derived from Table A-2. Check the results against the formulas in (4.8.5): $\mu = np$, $\sigma^2 = np(1-p)$.

4.8.5 Let Y be a random variable which follows a normal distribution with mean $\mu = 20$ and standard deviation $\sigma = 2$. What is the probability that a random observation taken from this population would be: (a) $\leqslant 21.0$, (b) $\leqslant 18.6$, (c) $\geqslant 23.2$, (d) between 17.4 and 22.4?

4.8.6 The average seasonal rainfall in Town A is 50 centimeters, with a standard deviation of 15 centimeters. Assume seasonal rainfall to be normally distributed. In a 75-year period, how many years would you expect to have between 38 and 65 centimeters of rain? How many drought years would you expect, if drought is defined as seasonal rainfall of 29 centimeters or less?

4.8.7 In a certain bakery process, the heights of cakes baked under standard conditions have a normal distribution with mean $\mu = 110$ millimeters and standard deviation $\sigma = 10$ millimeters.
   a. What is the probability that a cake from this population will be taller than 135 mm?
   b. If 10,000 cakes were baked, how many of them would you expect to be: (a) higher than 122 millimeters, (b) lower then 100 millimeters?

4.8.8 The acceptability of a capillary tube for a freezer is found by measuring the pressure drop in pounds per square inch between the two ends of the tube. The pressures obtained from a manufacturing process of capillary tubes show an average of 130 pounds per square inch and a standard deviation of 4 pounds per square inch. Assume that these pressures are random and normally distributed. Determine what: (a) percent of the pressures are below 121.6 pounds per square inch, (b) percent of the pressure readings lie between 121.6 and 134.4 pounds per square inch, (c) value is exceeded by 75 percent of the pressure readings, and (d) limits include the middle 90 percent of the pressure readings.

4.8.9 Suppose that in a certain population the true mean systolic blood pressure is 125 (millimeters of mercury) and the standard deviation of the distribution of individual blood pressures is 9. The distribution is considered to be normal.
   a. What proportion of the population has systolic blood pressure between 110 and 135?
   b. What percentage of the population has systolic blood pressure in the ranges 125 ± 9, 125 ± 18, and 125 ± 27, respectively?
   c. What proportion of the population has systolic blood pressure: higher than 145? as low as 100?
   d. If it is desired to classify in some special way (say Category C) the 5 percent of the population having the highest blood pressure, what is the criterion, in terms of blood-pressure reading, for putting an individual into Category C?

4.8.10 Suppose that health authorities wanted to impose a 95 percent effective quarantine on persons bitten by the *Anopheles* mosquito. (We assume a normal distribution of incubation periods with $\mu = 14$ days and $\sigma = 2$ days.) They want to determine times y_1 and y_2 after exposure such that only 2.5 percent of cases will develop malaria before y_1 and only 2.5 percent of cases will come down after y_2. Then the quarantine can extend from y_1 to y_2 days after exposure. Determine the quarantine period. Determine a 99.9 percent effective quarantine period.

4.9 NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

In our discussion of the binomial distribution in Sections 4.5 and 4.8, and in the examples considered there, it must have occurred to the reader that any reasonably large number n of trials would require either enormous tables or enormous calculations, or both, to produce the probabilities of various numbers of successes. This is in fact the case, and some of the earliest research in probability theory was directed to finding convenient yet close approximations to the exact probabilities.

The most successful of these approximation attempts led directly to the standard normal distribution. Recall the nature of the binomial distribution. The random variable Y can be looked on as the number of "successes" in n independent trials of an experiment wherein the probability of success in a single trial is p. As set forth in (4.8.5), the mean μ and standard deviation σ for this Y are:

$$\mu = np, \qquad \sigma = \sqrt{np(1-p)}.$$

The pertinent theorem states that the "standardized" or "normalized" version of Y, namely,

$$\frac{Y - \mu}{\sigma},$$

that is to say,

$$\frac{Y - np}{\sqrt{np(1-p)}}$$

has a distribution which becomes more and more nearly like the standard normal distribution as n gets larger and larger. We say that the fraction is *asymptotically normal*, meaning thereby that

$$\text{Distribution of } \frac{Y - np}{\sqrt{np(1-p)}} \to N(0, 1) \text{ as } n \to \infty. \tag{4.9.1}$$

In practical use, we stop far short of infinity and take $N(0, 1)$ as an *approximation* to the distribution of the indicated fraction. The goodness (closeness) of the approximation depends on both n and p. Experience has shown the approximation satisfactory in all cases where *both np and $n(1-p)$ are greater than 5*.

We summarize all the foregoing in the procedural rule:

> If Y is binomial with parameters n and p
> such that $np > 5$ and $n(1-p) > 5$, then
>
> $$\frac{Y - np}{\sqrt{np(1-p)}} \approx Z,$$
>
> where Z is standard normal $N(0, 1)$.

(4.9.2)

Example 4.9.1

A baseball player who has a .300 batting average would be "expected" to have six hits in 20 turns at the plate. Assuming actuality to satisfy the conditions for the binomial distribution (independent trials, constant hit probability), what is the probability that he will get no more than six hits in the 20 times at bat?

Taking Y = the number of hits, we have Y binomial with $n = 20$, $p = .30$, and we want $P(Y \leqslant 6)$. From Table A-2, the exact probability (to four decimal places) is 0.6080.

Since here $np = 20(.3) = 6$, $n(1-p) = 20(.7) = 14$, and the values 6 and 14 are both greater than 5, the approximation (4.9.2) is justified, and we could argue:

$$\mu = np = 20(.3) = 6, \qquad \sigma = \sqrt{np(1-p)} = \sqrt{20(.3)(.7)} = \sqrt{4.20} = 2.05;$$

$$P(Y \leqslant 6) = P\left(\frac{Y-6}{2.05} \leqslant \frac{6-6}{2.05}\right) \approx P\left(Z \leqslant \frac{0}{2.05}\right) = P(Z \leqslant 0) = .5000,$$

the final probability figure coming from the standard normal distribution (Table A-3).

So our approximation gives the answer .5000 to a question whose exact answer is .6080. Not terribly bad, considering the relatively small value of n and of np. But the approximation can be greatly improved by a small adjustment.

The adjustment made when applying the approximation (4.9.2) to a specific case is often termed a *correction for continuity*. Both the name and the method of making the correction are clarified by observing the motivation for the correction. The standard normal distribution is represented by a continuous curve, while the binomial distribution is represented by a histogram. Our approximation of a binomial probability is thus an approximation of an area made up of rectangles by an area under the continuous normal curve. Thus by reference to the diagram, we see the reasonableness of a "correction" to improve the approximation:
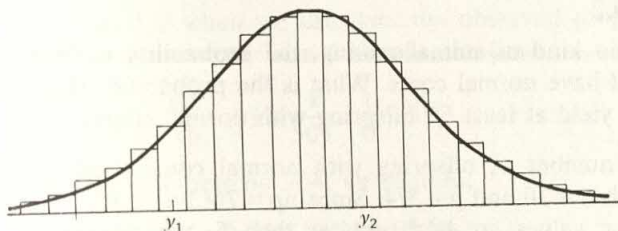
$$P(Y \leqslant y_1) \approx P\left(Z \leqslant \frac{(y_1+\frac{1}{2})-np}{\sqrt{np(1-p)}}\right),$$

$$P(Y \geqslant y_2) \approx P\left(Z \geqslant \frac{(y_2-\frac{1}{2})-np}{\sqrt{np(1-p)}}\right),$$

$$P(y_1 \leqslant Y \leqslant y_2) \approx P\left(\frac{(y_1-\frac{1}{2})-np}{\sqrt{np(1-p)}} \leqslant Z \leqslant \frac{(y_2+\frac{1}{2})-np}{\sqrt{np(1-p)}}\right).$$

(4.9.3)

The correction for continuity is the adjustment 1/2 made in the manner shown in the foregoing three statements.

Example 4.9.1a

We improve the approximation in our calculation of probability of no more than six hits by using a small rough sketch to show us how to adjust the calculation in accordance with the correction for continuity. We want to approximate that area of the histogram that is indicated by shading, and so we see that the area under the approximating curve should be taken to the boundary 6.5. Then we calculate:

$$P(Y \leqslant 6) = P\left(\frac{Y-6}{2.05} \leqslant \frac{6-6}{2.05}\right) \approx P\left(Z \leqslant \frac{6.5-6}{2.05}\right) = P\left(Z \leqslant \frac{0.5}{2.05}\right) = P(Z \leqslant 0.24).$$

Table A-3 shows the answer to between .5793 and .6000. (Taking $0.5/\sqrt{4.20} = 0.5/2.049 = .2440$ and using more elaborate standard normal tables, we would have the answer .5964.) This is now very close to the exact answer .6080.
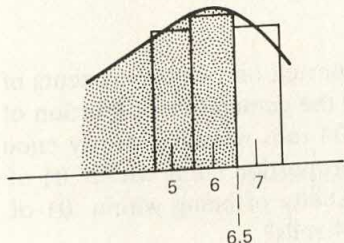
Example 4.9.2

In a certain kind of animal mating, the probability is 75 percent that the offspring will have normal coats. What is the probability that 76 independent matings will yield at least 50 offspring with normal coats?

Take $Y =$ number of offspring with normal coats, and consider Y to be binomial with $n = 76$ and $p = 3/4$. Since $np = 76(3/4) = 57$, $n(1-p) = 76(1/4) = 19$, and these values are both greater than 5, the approximation (4.9.2) is applicable.

$$\mu = np = 76(3/4) = 57,$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{76(3/4)(1/4)} = \sqrt{14.2} = 3.77;$$

$$P(Y \geqslant 50) = P\left(\frac{Y-57}{3.77} \geqslant \frac{50-57}{3.77}\right)$$

$$\approx P\left(Z \geqslant \frac{49.5-57}{3.77}\right) \quad \text{(See diagram below)}$$

$$= P\left(Z \geqslant \frac{-7.5}{3.77}\right) = P(Z \geqslant -1.99).$$

Taking one decimal place in y so as to match Table A-3, we have

$$P(Y \geqslant 50) \approx P(Z \geqslant -2.0) = 1 - P(Z < -2.0) = 1 - .0227 = .9773.$$

The exact probability, to four decimal places, as given in extensive tables of the binomial distribution, is .9735.
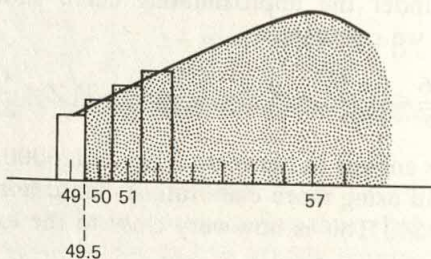
Example 4.9.3

In Section 4.3 we reported on two experiments of rolling a pair of dice 504 times, keeping track of the cumulative proportion of rolls that gave point "7." We commented that 504 rolls were not nearly enough to give high likelihood of ending up with the proportion being within .01 of the theoretical probability 1/6. What is the probability of being within .01 of 1/6 when we observe the proportion of 7s in 504 rolls?

To be within .01 of 1/6 when we calculate the observed proportion of 7s means that, if Y = the number of 7s in 504 rolls, we have

$$\frac{1}{6} - .01 \leqslant \frac{Y}{504} \leqslant \frac{1}{6} + .01,$$

that is, multiplying all members by 504, we require:

$$84 - 5.04 \leqslant Y \leqslant 84 + 5.04,$$
$$78.96 \leqslant Y \leqslant 89.04.$$

Since the number of 7s has to be a whole number, the boundaries on Y consistent with the above are

$$79 \leqslant Y \leqslant 89.$$

Taking Y as binomial with $n = 504$ and $p = 1/6$, we have

$$\mu = np = 504(1/6) = 84,$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{504(1/6)(5/6)} = \sqrt{70.0} = 8.37;$$

$$P(79 \leqslant Y \leqslant 89) = P\left(\frac{79-84}{8.37} \leqslant \frac{Y-84}{8.37} \leqslant \frac{89-84}{8.37}\right)$$

$$\approx P\left(\frac{78.5-84}{8.37} \leqslant Z \leqslant \frac{89.5-84}{8.37}\right)$$

$$= P\left(\frac{-5.5}{8.37} \leqslant Z \leqslant \frac{5.5}{8.37}\right)$$

$$= P(-0.7 \leqslant Z \leqslant +0.7) = .7580 - .2420 = .5160.$$
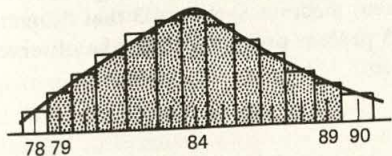
EXERCISES

4.9.1 In the case of certain animals, the probability that an offspring is pure white is 25 percent. What is the probability that among 300 offspring, the number of pure white will be: (a) no more than 67, (b) more than 90?

4.9.2 According to the latest published mortality tables, about 23 percent of persons 60 years of age die before reaching age 70. What is the probability that of a group of 2000 persons of age 60 the number of deaths before 70 will be at least 450 and not more than 500? What would you say about the group if the number of deaths actually turned out to be 420?

4.9.3 If a type of surgical operation having a 90 percent probability of success is performed 100 times, what is the probability that there will be more than 11 failures?

4.9.4 It is believed that 20 percent of the voters in a certain community are Independent voters. A poll is taken of 400 voters constituting a random sample. Of these, 95 state that they are Independent. What is the probability of a response as large as this if the 20 percent proportion is correct?

4.9.5 With the standard process of manufacturing a certain article, 2 percent of the produced units are defective. A new time- and money-saving process will be installed if it does not significantly increase this proportion of defectives. A test run of 2500 units produced by the new process shows 55 defective. What is the probability of a number of defectives as large as this if the 2 percent defective rate has been maintained?

4.9.6 The Mecca College basketball team has a long-run record of 60 percent wins. What is the probability that it will not do better than break even in next season's 24 games?

4.9.7 Repeat the argument of Example 4.9.3 using 5400 rolls of the dice, and thus confirm the statement made in Section 4.3 that "something over 5000 rolls are required to give 95 percent probability that the observed proportion (of 7s) will be within .01 of 1/6."

5.1 THE USE OF A RANDOM SAMPLE

When we use the data of a sample to tell us something about the population from which the sample was taken, the *only* thing we can be *sure* of is that we do *not* have the *whole* truth about that population. A frequency distribution or a histogram of the y values in the sample can give us a good idea about the probability distribution of the characteristic Y in the population, but it cannot give us the exact distribution of Y. The mean $\bar{y}$ of the sample gives us an *estimate* of the population's mean μ, but $\bar{y}$ can never give us μ precisely. The same is true of the sample standard deviation s: s is *never* σ; it is an *estimate* of σ.

In such a state of logical ignorance, the best we can do is to hedge our estimates in some scientific manner that takes chance into account. Then at least we can have a high degree of confidence that the truth lies within certain specified limits. This is the purpose of that part of statistical inference called *statistical estimation*. It is one reason that statistical inference is sometimes defined as *decision-making in the face of uncertainty.*

When a gambler places a bet on the roll of a pair of dice or on the outcome of a horse race, he is making a decision based on the odds assumed to hold for the various outcomes—a judgment based on the probability distribution of the chance process. So it is with a statistician making an estimate

5

Educated Guessing

5

about a population; he makes a decision based on the probability pattern assumed to govern sample estimates.

In one important sense the statistician is worse off than the gambler. After the dice are rolled or after the horse race is run, the gambler knows exactly whether his decision was right or wrong. In most cases the statistician can never *know* whether his hedged estimate is right or wrong. Even if he estimates voter preferences, the actual outcome of the election will not tell him whether he was right or wrong *at the time of his sampling*. He must rely on an estimation procedure that has *a very high probability* of giving a correct bracket each time it is used with a single sample.

Mathematicians who specialize in the probability theory underlying samples have worked out the probability patterns governing estimates in a wide variety of situations. In all cases of practical usefulness, these patterns are based on having the sample be a *random sample*.

We present a method for drawing a sample which in most practical situations will satisfy the requirements. The important thing to keep in mind is that our methods of educated guessing about the limits to put around an estimate will be valid *only* if the estimate comes from a *random sample*.

5.2 DEFINITION OF A RANDOM SAMPLE

In order to get an observation y on the random variable Y, we operate the chance process that governs the distribution of Y and observe the result. We roll the dice and observe the resulting point. We toss the coins and observe the number of heads. We inspect a television tube taken from a production line and see whether it is defective or nondefective. We take a capsule from the military draft lottery bowl of birthdates and read the enclosed date. We ask a laboratory to analyze a vial of blood and discover the cholesterol content. We select a student file and read the G.P.A.

If the observation is to be a *random* observation, its outcome must be determined solely by the probability distribution of Y. And this means that the operation of the chance process—the *trial* of the random experiment—must be free from any influence outside the probability pattern of Y. In operational terms, this means that the single trial that we conduct must be just as likely as any other trial that we could conduct.

A *fair* roll of a pair of dice or a *fair* toss of three coins can reasonably be assumed to meet the above requirement. For a production line of television tubes, the bowl of birthdate capsules in the draft lottery, a collection of vials of blood, or a cabinet of student files, we satisfy the requirement by using some selection procedure that makes each element (tube, capsule, vial, or file) as likely to be drawn as every other element.

A random observation made in such a manner is a random variable: its value is unknown before the trial is made, and the value that appears in the trial is determined by the probability distribution of the population random variable Y. The random observation *is* a Y. If it is our first observation in a series of observations, we can show this by subscripting Y_1. After the trial we know the outcome and call it y_1.

If a second observation is made in the same manner as the first, it is another Y random variable; we can call it Y_2. Y_1 and Y_2 are called *independent* observations if the outcomes of the first and second trials are both independent events in the sense mentioned in Chapter 4.

We now extend the process to the third, fourth, . . . , nth observations. In this way we have n independent random observations Y_1, Y_2, . . . , Y_n, each governed by the probability distribution of Y. The outcomes of the n independent trials are the observed numerical values $y_1, y_2, \ldots, y_n$. Such a set of random observations is called a *random sample* of size n from the population of Y.

DEFINITION 5.2.1 A random sample of size n from the population of Y is a set of n independent random variables Y_1, Y_2, . . . , Y_n, each having the probability distribution of Y.

All the foregoing discussion presupposes that the number of possible observations is theoretically unlimited. This is true of dice rolls or coin tosses. It cannot be true if the trial involves a distinct member of a finite collection. Television tubes, draft-lottery capsules, vials of blood, and student files are not infinite in number. Here we get into complications that require advanced study in sampling. In such situations one talks about a *random sample from a finite population* and defines that as a set of n elements drawn from the finite population in such a way that each possible set of n is as likely as every other possible set of n. When n is a very small fraction of the total number of available elements, the sample behaves in a manner very close to that of our definition above. *In this book we shall limit ourselves to such cases.*

5.3 DRAWING A RANDOM SAMPLE FROM A FINITE POPULATION

Our intuition gives us a reasonably good idea of how to draw n items at random from a collection of items; label the items in some identifiable manner, put the labels in a box, shake the box energetically, and then draw out n labels "at random." But those last two words are more easily said than done. The hand grabbing can have a bias for the top of the pile, the bottom of the pile, this or that side of the box, or a pattern of spacing grabs. The more consciously we try to work out a system of random selections, the less likely it is to be random.

Scientists realized long ago that it would be preferable to have a table of random numbers assembled from the most refined hand grabbing possible and tested for absence of patterns or other bias. Such a table could be an extremely long sequence of digits, each of which is 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9, and each of which was the result of a random drawing wherein the 10 possibilities were equally likely. We could print this sequence of digits in columnar form, column after column, and on as many pages as necessary. Then any random start in the table would give us a random sequence of digits from there on. We could read the columns two at a time to give us two-digit random numbers, three at a time for three-digit random numbers, and so on.

A small-scale procedure for producing a random sequence of digits could be the following. In each of 100 capsules place a slip of paper marked 0, in each of 100 other capsules place a slip marked 1, in each of another 100 capsules place a slip marked 2, and so on by batches of 100 capsules until we have 1000 capsules representing the digits 0–9 in equal proportions. Put the capsules in a large box, shake the box vigorously, and draw out a capsule. Record the digit in the capsule, replace the capsule in the box, shake the box again, and draw out a capsule. Record *that* digit, replace the capsule, shake the box, and draw out a capsule. Continue on and on as long as interest and strength allow.

The earliest published tables of random digits were constructed by some variation of the above physical process. After the invention of high-speed electronic computing machines, programs were devised to command such a computer to produce sequences of random digits. This is now the standard procedure for producing sequences of random digits. The best known of published tables is *A Million Random Digits*, authored by the RAND Corporation (Santa Monica, California) and published by the Free Press (Glencoe, Illinois) in 1955. The volume is often referred to as just the *RAND Table*. The sequence of 1,000,000 random digits is printed in 20,000 rows of 50 digits each, fifty rows to a page for 400 pages. One page of this table (chosen at random!) is reproduced as our Table A-6 in the Appendix.

Rows and columns are printed in blocks of five for ease in reading. Row numbers are given at the left; column numbers must be counted off by the reader. Our Table A-6 is seen to contain rows 03000–03049 of the RAND Table; these appear on the RAND Table's page 61, the randomly chosen page. To check the numbering of columns, take row 03034 as example and note that the digit in column 1 is 8, the digit in column 2 is 1, the digit in column 3 is 8, the digit in column 4 is 0, and so on; the digit in column 37, for example, is 9.

To use a table of random digits, one must take a random starting point in the table and then read a sequence from there. One may read across rows, going forward or backward; he may read along columns, down or up. To get a sequence of random *numbers* he may take the digits in blocks of two, three, four, and so on—as many digits as required to fit the largest item number he has to accommodate. For example, if the largest number in the set from which we are going to draw is 27,654, then we are going to have to read digits in batches of five, so as to allow us to get numbers up through 27,654. Numbers with fewer digits will appear like 00218, 00006, and so on.

There are various devices for choosing the random starting place. In Table A-6 a reasonable procedure is as follows. There are in this table 50 rows (row numbers ending 00–49) and 50 columns (we would count them off 1, 2, ..., 50). So we have to choose at random a pair of numbers to determine our starting point, one number between 00 and 49 to fix the row, and one number between 1 and 50 to fix the column. With eyes closed, let a finger pick a point on the page. With eyes opened (!) take the five-digit block of numbers nearest your finger as a *code number* for the starting place. For example, consider what happened to one of the authors: his finger came nearest the block 57429 (it is the fifth block in row 03012). We take the first two digits of the five-digit block to tell us a row number, reducing by 50 if necessary to get a number between 00 and 49; in our example this gives us row 07 (57 minus 50). We take the next two digits in the block to give us a column number, again reducing by 50 if necessary to get a number between 01 and 50; in our example we get column 42 directly. The resulting row and column numbers then specify the starting place for our random sequence. In our example we have row 07, column 42. The reader should verify that the resulting sequence, if one reads down the column, starts out 4, 0, 9, 4, 4, 2, 3, 6, ...; and if one reads left to right along a row, the sequence starts out 4, 2, 6, 8, 9, 2, 9, 5, 0, ... We pursue a sequence from one column to the next if we are reading by columns and from one row to the next if we are reading by rows. Thus the row-wise sequence above would be extended to read 4, 2, 6, 8, 9, 2, 9, 5, 0, 7, 2, 8, 9, 2, ...

Let us now use this procedure to select a random sample of nine students from the 180 students considered in Chapter 1. The students are individually

identified by number, 1–180. We must therefore choose at random nine numbers from the collection {1, 2, 3, ..., 180}. This means that we must use random digits in blocks of three, so that numbers up through 180 can appear. There will of course be many random numbers which we will have to pass over—any three-digit block giving a number larger than 180. We shall also pass over any random number that we have already used.

Again permit the authors to finger the starting point code to set up the example. This time the code block turned out to be 40746 (eighth block in row 03001). This gives us the start at row 40 and column 24 (74 minus 50). Reading down a column seems to us the easiest way to read figures, and so we shall proceed that way. From the starting digit the sequence looks as follows on the left, and is read as shown to the right:

98	0	980
83	5	835
31	0	310
81	8	818
17	2	172
08	9	089
81	1	811
66	1	661
77	1	771
15	8	158

Up to this point we have produced three usable numbers: 172, 089, and 158. We must continue on, starting at the top of the *next* column. There we read 783, 149, 584, and so on. The complete sequence is as follows, with usable numbers starred.

980	*158	458	465	328
835	783	960	598	458
310	*149	702	235	370
818	584	300	980	552
*172	282	288	356	670
*089	401	686	*156	589
811	*120	342	289	*074
661	339	*017	987	
771	266	548	*109	

Thus our random sample of nine students is composed of students number 172, 89, 158, 149, 120, 17, 156, 109, and 74.

Suppose our interest is in the G.P.A. We make the observations from the information in Chapter 1, and then can set down our sample as follows (with the y^2 column added for calculation of s^2):

Student Number	G.P.A. y	y^2
172	1.00	1.0000
89	3.36	11.2896
158	2.17	4.7089
149	1.93	3.7249
120	3.19	10.1761
17	1.13	1.2769
156	2.91	8.4681
109	2.73	7.4529
74	2.40	5.7600
	20.82	53.8574

The reader should confirm that this sample gives

$$\bar{y} = \frac{20.82}{9} = 2.313; \qquad s^2 = \frac{53.8574 - \frac{(20.82)^2}{9}}{8} = 0.7117, \qquad s = 0.844.$$

The question now is: How close to the *population* mean μ and standard deviation σ can we believe these values to be?

EXERCISES

5.3.1 In Table A-6, confirm that the method described in the text for choosing a random starting place in the table will give the first number to be read as shown in the following examples.

	Starting Point Code Number	Number of Digits per Block	First Number to be Read
(a)	04187	3	677
(b)	28351	2	27
(c)	72224	3	358
(d)	94452	4	5202
(e)	13700	3	229
(f)	02831	4	8127
(g)	85790	3	065
(h)	50662	2	81

5.3.2 Using Table A-6, draw sets of random item numbers as specified in the following examples.

	Item Numbers in Total Set	Required Size of Sample	Starting Point Code Number
(a)	1, 2, 3, ..., 873.	6	30170
(b)	1, 2, 3, ..., 4618.	5	69033
(c)	1, 2, 3, ..., 659.	7	47951
(d)	1201, 1202, 1203, ..., 7890	8	56704

5.3.3 Draw a random sample of 16 students from the Mecca Community College class (Table 1.2.1). Set down the data on these 16 students, and give descriptive statistics as follows: (a) proportion female, (b) proportion having no party preference, (c) proportion in favor of legalizing marijuana, (d) mean scale score on the marijuana question, (e) mean commuting distance, (f) mean G.P.A. average, and (g) G.P.A. standard deviation. An interesting survey can be made by comparing your results with those of the other members of your class.

5.3.4 A listing of 59 Accounts Receivable ledgers is shown in Table 5.3.1. The data are the total dollar amounts, rounded to the nearest thousand of dollars, contained in each of the ledgers.

a. Construct a frequency distribution of these accounts, using intervals of one thousand dollars starting at 6,500.

b. Construct a cumulative percentage frequency distribution. Using this distribution, give the median dollar amount in the 59 ledgers.

c. Calculate the mean dollar amount in the 59 ledgers.

d. Using a table of random numbers, select two samples of size $n = 9$ each. For each sample, calculate its mean.

e. When the exercise is finished, pool your two sample means with those of the other class members and plot them on the same scale as used in (a) above.

f. Compare the distribution in (a) with that found in (e).

TABLE 5.3.1 Listing of Accounts Receivable
Ledgers in Ledger Number Order

1	11,000	21	13,000	41	11,000
2	14,000	22	9,000	42	8,000
3	10,000	23	12,000	43	11,000
4	11,000	24	11,000	44	14,000
5	10,000	25	8,000	45	11,000
6	16,000	26	8,000	46	9,000
7	9,000	27	9,000	47	10,000
8	11,000	28	10,000	48	12,000
9	10,000	29	10,000	49	12,000
10	9,000	30	8,000	50	12,000
11	9,000	31	10,000	51	9,000
12	10,000	32	11,000	52	9,000
13	11,000	33	10,000	53	10,000
14	9,000	34	11,000	54	8,000
15	13,000	35	9,000	55	9,000
16	11,000	36	9,000	56	9,000
17	8,000	37	11,000	57	7,000
18	9,000	38	11,000	58	10,000
19	11,000	39	12,000	59	10,000
20	8,000	40	10,000		

5.4 THE PROBABILITY DISTRIBUTION OF A SAMPLE MEAN

Since the observations in a random sample are random variables, the mean of those observations is a random variable. With our capital-letter convention, we can write

$$\bar{Y} = \frac{\sum Y}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$

When we draw a random sample of size n from the population of Y, we then obtain the observed values $y_1, y_2, \ldots, y_n$. Calculating the mean of these values gives us the value of $\bar{y}$ for the sample mean. This value of $\bar{y}$ is thus the outcome of one observation on the random variable $\bar{Y}$. Take another random sample of size n from the population of Y and we shall end up with a $\bar{y}$ value different from the first. And so on. That is, there is a population of $\bar{y}$-values just as there is a population of y-values. Thus we speak of the *probability distribution of the sample mean* $\bar{Y}$.

With somewhat more mathematics than we want to use in this book, we could work out various important facts concerning the probability distribution of $\bar{Y}$. With still more advanced mathematical methods we could derive one of the fundamental laws on which much of statistical inference is based. Here our interest is in *applying* such facts and laws, and so we shall simply state and use

them, leaving derivations for the reader to consider if someday he makes a more detailed study of statistics.

If Y has mean μ and standard deviation σ, then, no matter what is the specific form of the probability distribution of Y, the probability distribution of $\bar{Y}$ has the following mean and standard deviation:

$$\begin{cases} \text{Mean of } \bar{Y}: & \mu_{\bar{Y}} = \mu \\ \text{Standard deviation of } \bar{Y}: & \sigma_{\bar{Y}} = \sigma/\sqrt{n} \end{cases} \qquad (5.4.1)$$

Here we see two of the basic facts that make the sample mean so important in statistics: the mean of the probability distribution of $\bar{Y}$ is precisely the same as the mean of the population of Y, and the standard deviation in the distribution of $\bar{Y}$ is *smaller* than that in the population of Y as soon as the sample size n is larger than 1. This means that $\bar{Y}$ varies about an expected value that is the same as the expected value of Y, and its variation is smaller than the variation in Y. Moreover this variation in $\bar{Y}$ becomes less and less as the sample size is made larger and larger. This makes intuitive sense, for we would anticipate that an infinitely large sample should give a $\bar{y}$ that would hit μ on the nose, and this is indicated by the fact that $\sigma/\sqrt{n}$, measuring the variation in $\bar{Y}$, goes to zero as n goes to infinity.

Information about the *form* of the probability distribution of $\bar{Y}$ is contained in the following two important theorems.

Theorem 5.4.1

If Y has the *normal* distribution with mean μ and standard deviation σ, then $\bar{Y}$ has a *normal* distribution, with mean μ and standard deviation $(\sigma/\sqrt{n})$. In shorthand notation,

$$\text{if } Y \text{ is } N(\mu, \sigma), \text{ then } \bar{Y} \text{ is } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

The mean and standard deviation of $\bar{Y}$ in this theorem are simply in accordance with (5.4.1); the important new fact is that $\bar{Y}$ has a normal distribution if Y has.

The next theorem is one of the most important, and remarkable, laws in probability theory. It shows that the distribution of $\bar{Y}$, as n is made larger and larger, has a tendency to become more and more nearly a *normal* distribution *no matter what nonnormal distribution Y has*. This law is called the *central limit theorem*.

Theorem 5.4.2 (Central Limit Theorem).

If Y has a nonnormal distribution with mean μ and standard deviation σ, then, as n increases without bound, the distribution of $\bar{Y}$ approaches the normal distribution with mean μ and standard deviation $(\sigma/\sqrt{n})$.

Using the formula for transforming a normal random variable into the standard normal random variable Z, we can summarize the above facts in the following operational rule:

$$\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \begin{cases} = Z & \text{if Y is normal,} \\ \approx Z & \text{if Y is nonnormal and } n \text{ is large,} \end{cases} \tag{5.4.2}$$

where Z has the standard normal distribution $N(0, 1)$.

How large is "large" for n? Much effort has gone into the study of the closeness of the approximation for various values of n. There are a number of rules of thumb but no universally held standards, since so much depends on the specific form of the distribution of Y and on the criterion of "closeness" which one adopts for assessing the approximation. Figure 5.4.1 indicates how rapidly the distribution of $\bar{Y}$ approaches the normal; it suggests that $n \geq 30$ is large enough to bring the distribution of $\bar{Y}$ close to normal.
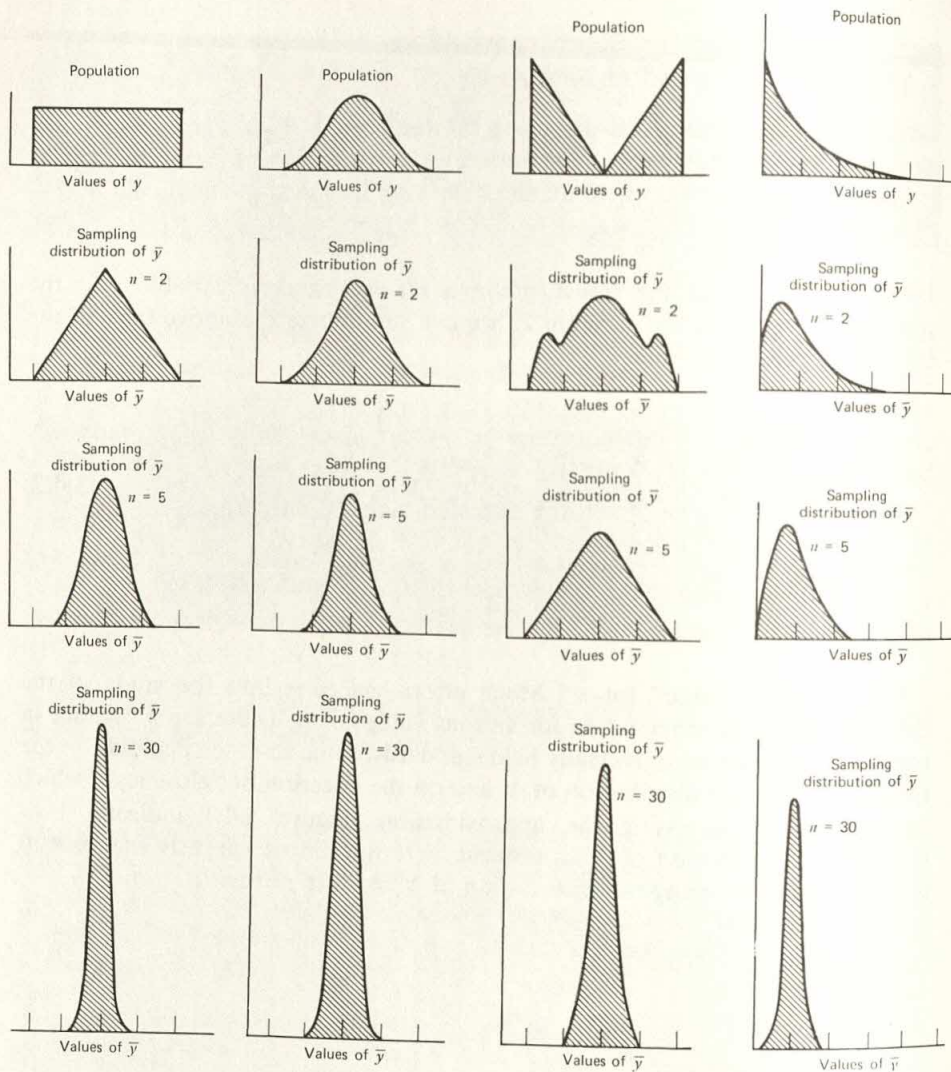
FIGURE 5.4.1 Distribution of the sample mean for various populations and sample sizes. Adapted with permission from Kurnow et al., *Statistics for Business Decisions*, Richard D. Irwin, Inc., Homewood, Illinois, (1959c), pp. 182–183.

5.5 A CONFIDENCE INTERVAL FOR μ WHEN σ IS KNOWN

The relation (5.4.2) is a statement about the deviation of $\bar{Y}$ from the population mean μ. It shows that if we know σ we can study the probability behavior of the deviation $\bar{Y} - \mu$ by using the standard normal distribution of Z.

When an investigator uses $\bar{Y}$ as an estimator of μ, he would like to have some bound on the magnitude of the deviation that has a high probability of not being exceeded. He could hope, for example, that there is 95 percent probability that the deviation will not exceed a certain reasonable amount. If he gets that level of assurance, he could then have 95 percent *confidence* that the $\bar{y}$ which actually occurs in a sample is within the specified distance from μ.

The distribution of Z (Table A-3) tells us that the central 95 percent of the probability of z-values is bounded by $z = -1.96$ and $z = +1.96$.
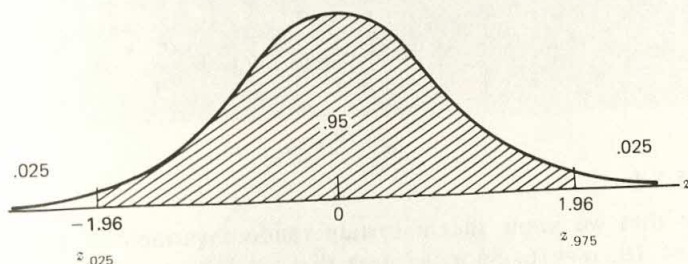
This z interval we can now translate into an interval concerning $\bar{Y}$ and μ, by using (5.4.2) and a bit of algebra. We pursue our probability statement as follows.

$$.95 = P(-1.96 < Z < 1.96)$$

$$= P\left(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$= P\left(-1.96 \frac{\sigma}{\sqrt{n}} < \bar{Y} - \mu < 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(-\bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\bar{Y} + 1.96 \frac{\sigma}{\sqrt{n}} > \mu > \bar{Y} - 1.96 \frac{\sigma}{\sqrt{n}}\right),$$

that is,

$$.95 = P\left(\bar{Y} - 1.96\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + 1.96\,\frac{\sigma}{\sqrt{n}}\right). \tag{5.5.1}$$

This statement tells us that there is 95 percent probability that $\bar{Y} -$ $1.96(\sigma/\sqrt{n})$ and $\bar{Y} + 1.96(\sigma/\sqrt{n})$ will *bracket* the population mean μ. This means that in the long run of repeatedly taking random samples of size n from the population of Y, 95 percent of the time the sample will give a $\bar{y}$ such that $\bar{y} - 1.96(\sigma/\sqrt{n})$ and $\bar{y} + 1.96(\sigma/\sqrt{n})$ bracket the mean μ. Thus when we draw just a *single* sample, we are willing to bet 95:5 that *its* $\bar{y}$ is the kind that gives a bracket of μ by $\bar{y} \pm 1.96(\sigma/\sqrt{n})$. We express this by saying that we have 95 percent *confidence* in stating that μ lies between $\bar{y} - 1.96(\sigma/\sqrt{n})$ and $\bar{y} + 1.96(\sigma/\sqrt{n})$. We call the bracket interval a 95 percent *confidence interval*.

A 95 percent confidence interval for the population mean μ is:

$$\bar{y} - 1.96\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + 1.96\,\frac{\sigma}{\sqrt{n}}.$$

Example 5.5.1

Suppose that we know that a certain random variable Y has a standard deviation of 16: $\sigma = 16$. Suppose next that we take a random sample of 64 observations on Y, calculate the mean of the resulting values, and find that to be 142.61: $\bar{y} = 142.61$. We can now proceed as follows.

The limits of a 95 percent confidence interval for μ are

$$142.61 \pm 1.96\left(\frac{16}{\sqrt{64}}\right) = 142.61 \pm 1.96\left(\frac{16}{8}\right)$$

$$= 142.61 \pm 1.96(2)$$

$$= 142.61 \pm 3.92.$$

Then by using the minus and the plus values, we get $142.61 - 3.92 = 138.69$ as the lower limit, and $142.61 + 3.92 = 146.53$ as the upper limit, so that we can state that a 95 percent confidence interval for the population mean μ is $138.69 < \mu < 146.53$.

This tells us the limits $(138.69, 146.53)$ between which we believe μ lies, and the degree of confidence (95 percent) that we can have in the procedure that gave us the limits.

We can of course arrange for any degree of confidence other than 95 percent by simply changing the z value from 1.96 to that value in Table A-3 which gives us the desired area under the standard normal curve. We can state the general confidence interval as follows.

Given a random sample of n observations from the population of Y, a c percent confidence interval for the population mean μ is

$$\bar{y} - z_*\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{y} + z_*\left(\frac{\sigma}{\sqrt{n}}\right), \tag{5.5.2}$$

where $\bar{y}$ = the mean of the sample;

σ = the standard deviation of the population of Y;

n = the number of observations in the sample;

z_* = the confidence-limit value of z, determined by

$$P(-z_* < Z < z_*) = c \text{ percent.}$$
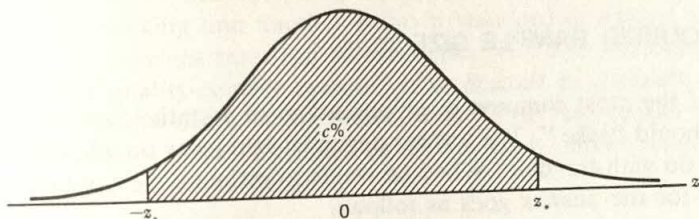
Example 5.5.2

Suppose that in the situation of Example 5.5.1 we want to have a 99 percent confidence interval for μ. With .99 area in the center of the z distribution, there is area .005 left out in the left tail and area .005 left out in the right tail, so that, according to Table A-3,

$$-z_* = z_{.005} = -2.58 \quad \text{and} \quad z_* = z_{.995} = +2.58.$$

Then the limits of a 99 percent confidence interval for μ are

$$142.61 \pm 2.58\left(\frac{16}{\sqrt{64}}\right) = 142.61 \pm 2.58(2)$$

$$= 142.61 \pm 5.16,$$

and our 99 percent confidence interval for μ is $137.45 < \mu < 147.77$.

Notice that this interval is *wider* than the 95 percent confidence interval in Example 5.5.1. This matches common sense, since we should expect to be forced to enlarge the interval around $\bar{y}$ if we want to increase our confidence in the bracket. Conversely we can shorten the confidence interval by reducing our degree of confidence.

The standard deviation of the sample mean $\bar{Y}$ has been given the special name *standard error of the (sample) mean*:

Standard error of the (sample) mean = standard deviation of $\bar{Y}$

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}.$$

With this nomenclature, one may find it useful to look on the limits in (5.5.2) as "$\bar{y}$ plus–minus so many standard errors." The "so many" is determined from the tables of Z to give the degree of confidence we want.

5.6 REQUIRED SAMPLE SIZE

One of the most common questions asked of a statistician is "How large a sample should I take?" The answer depends right away on what the questioner wants to do with the data. When he wants to estimate the population mean μ, the logic for the answer goes as follows.

In (5.5.2) we see $z_*(\sigma/\sqrt{n})$ as the allowance to place around $\bar{y}$ in order to give us a confidence interval for μ. We can then have c percent confidence that μ lies within that distance from $\bar{y}$. If we specify how close we want the estimate to be, we are specifying the value of $z_*(\sigma/\sqrt{n})$, and we have the beginning of an equation to solve for n. Let us designate the specified distance by d. Then our requirement for sample size is

$$\frac{z_* \sigma}{\sqrt{n}} \leq d. \tag{5.6.1}$$

With d specified, we still must know z_* and σ in order to solve for n. This is why the statistician asks for answers to the following questions before he ventures an opinion as to what sample size is needed.

a. How close to μ do you want to be? (The answer is d.)

b. How sure do you want to be that you are that close? (The answer gives c percent, and from that will give z_*.)

c. What can you tell me about the variability in the characteristic which you are measuring? (The answer is needed for deciding on a value for σ.)

The last question is of course the most difficult to answer. Past experience with similar studies may provide reliable information about σ. All else failing, the statistician may ask "What are the smallest and largest values ever observed for Y?" Here he has in mind the picture of the normal distribution, where virtually all (99.74 percent) of observations fall within ±3 standard deviations of the mean. That suggests a spread of 6σ between the smallest and largest predictable observations. He could then in his desperation take σ as one-sixth of the distance between the smallest and largest y-values ever observed.

Example 5.6.1

The Hoboken plant is producing 1-pound cans of Maxban Cottage Coffee. The net weight of coffee on the packing line has been set by Manufacturing Standards as $\mu = 16.04$ ounces. Historically, the weekly net weight records have shown that packing line number 6 has a standard deviation of $\sigma = .03$ ounce (i.e., the net weight from can to can varies with a $\sigma = .03$ ounce). For management and quality-control reasons, it is desired to estimate the true weekly average packed weight ±.01 ounce, with 95 percent confidence. How many 1-pound cans of coffee will have to be randomly chosen each week in order to guarantee these specifications?

n = sample size = ?

$\sigma = .03$

$z_* = 1.96$ (normal deviate from 5 percent tail area split 2.5 percent in each tail)

$d = .01$

Therefore

$$\frac{z_* \sigma}{\sqrt{n}} \leqslant .01$$

$$\frac{(1.96)(.03)}{\sqrt{n}} \leqslant .01$$

$$\frac{(1.96)(.03)}{(.01)} \leqslant \sqrt{n}$$

$$5.88 \leqslant \sqrt{n}$$

$$(5.88)^2 \leqslant n$$

$$34.57 \leqslant n$$

Thus $n = 35$ sample cans will be taken at random throughout each week on line 6, the net weight of coffee determined for each can, and the average of all 35 cans calculated. On the assumption of a normally distributed net coffee weight for 1-pound cans, the true mean weekly net weight will be estimated within $\pm.01$ ounce with this sample of $n = 35$ cans.

Example 5.6.2

Consider again the situation in Examples 5.5.1 and 5.5.2. There, with $\sigma = 16$ and $n = 64$, we can be 95 percent confident that μ is within ± 3.92 of $\bar{y}$, and 99 per cent confident that μ is within ± 5.16 of $\bar{y}$. Suppose that we want to be 99 percent confident that μ is within ± 3 of $\bar{y}$. Then we have

$$d = 3; \qquad z_* = 2.58; \qquad \sigma = 16;$$

and (5.6.1) gives

$$\frac{(2.58)(16)}{\sqrt{n}} \leqslant 3,$$

$$\frac{(2.58)(16)}{3} \leqslant \sqrt{n},$$

that is,

$$\sqrt{n} \geqslant 13.76,$$

$$n \geqslant 189.3376.$$

Since the sample size must be a whole number, we see that the smallest sample size that will accomplish what we want is $n = 190$.

It is worth noting the effect of each of the three conditions d, z_*, and σ on the sample size requirement. We can see this in general form if we solve the inequality (5.6.1) to give a general formula for n:

$$\frac{z_* \sigma}{\sqrt{n}} \leqslant d,$$

$$z_* \sigma \leqslant d\sqrt{n}$$

giving

$$n \geqslant \left[\frac{z_* \sigma}{d}\right]^2. \tag{5.6.1'}$$

Here we can see that the required sample size: (a) increases as the desired degree of confidence increases (since increasing confidence demands increasing z_*), (b) increases as the standard deviation increases, and (c) decreases as the allowable tolerance d increases. Moreover (5.6.1') shows that any of these effects is in proportion to the *square* of the determining factor.

EXERCISES

5.6.1 Measurement errors in laboratory tests are usually normally distributed. Suppose a certain laboratory test has a $\sigma = 1.4$.
  a. Find: (i) a 95 percent confidence interval for the true value of the determination using a test that comes out 17.5, (ii) a 99 percent confidence interval, and (iii) the lengths of these intervals.
  b. Suppose now that the experiment is run five times and the mean is used to find the confidence interval. What will the lengths be now?
  c. What would the lengths be if 50 tests were done?
  d. What is the fewest number of tests that you would have to run so that a 99 percent confidence interval would have length 2?

5.6.2 A dentist wants to obtain a 95 percent confidence interval estimate of the average pain threshold of patients as measured by a dolorimeter. The measurement is the amount of heat (in millicalories) to which the patient reacts. On the basis of other research work he feels that the results will be normally distributed with a standard deviation equal to 50 millicalories.
  a. He tests 100 patients and the resulting mean is $\bar{y} = 230$ millicalories.
    (i) Compute the 95 percent confidence interval for the true mean of the population.
    (ii) What is the interpretation of this confidence interval?
    (iii) Do you think that a second sample of 100 patients would yield a confidence interval of 190 to 210 millicalories? Why or why not?
  b. How many patients should the dentist study in order to estimate the true pain threshold of patients within 10 millicalories with 90 percent confidence?

5.6.3 Resort community A in the mountains has been claiming that its lake is the purest (unpolluted) in the whole state. Their claim is that the true mean total solids in milligrams/liter is 40 with a $\sigma = 5$. Resort community B took 100 samples from its lake and found mean total solids in milligrams/liter to be 39. Assuming σ to be the same for lake B, place 95 percent confidence limits on the true mean total solids in community B's lake. Do you think it should publicize its finding and claim an even purer lake? Why or why not?

5.6.4 The blood-clotting time of hemophiliacs is normally distributed with a mean of 5 minutes and a standard deviation of 2 minutes. A new drug has been marketed whose efficacy is based on reducing the blood-clotting time of hemophiliacs. A random sample of nine hemophiliacs was chosen and given the new drug; the average blood-clotting time was 4 minutes.
  a. Assuming that the standard deviation has remained the same, place 90 percent confidence limits on the true mean blood-clotting time of hemophiliacs using the new drug.
  b. How many hemophiliacs would have to be tested (assuming $\sigma = 2$) so that a 90 percent confidence interval would have length 0.5 minute?
  c. Would the 90 percent confidence interval in (b) guarantee that the new drug would shorten clotting time?

5.6.5 In a certain normal population σ is known to be 25. If it is required that one be 95 percent confident that the $\bar{y}$ of a sample from this population shall be within 4 of the true population mean μ, how large a sample must be taken?

5.6.6 Suppose it is known from long experience that the variability in a certain method of determining the concentration of a chemical in solution is given by $\sigma = .005$ (grams per cubic centimeter). Determine the number of measurements required to give a 99 percent confidence interval for concentration which is .001 grams per cubic centimeter wide.

5.6.7 Suppose that you want to take a sample for estimating μ, of sufficient size to give 95 percent confidence that the resulting $\bar{y}$ will be within 4 of μ. You decide that 18 is a "surely large enough" value to assume for σ. How large should the sample be?

5.6.8 In people classified as normal in health, the mean serum haptoglobulin is known to be 100 milligrams per 100 milliliters with a standard deviation of 40 milligrams per 100 milliliters. A random sample of 25 cancer patients were found to have a mean serum haptoglobulin of 114 milligrams per 100 milliliters. Using the known standard deviation, does a 95 percent confidence interval for mean serum haptoglobulin of cancer patients include the normal value of 100 milligrams per 100 milliliters?

5.7 A CONFIDENCE INTERVAL FOR μ WHEN σ IS UNKNOWN

In most practical situations σ is really just as much unknown as μ. We have to behave as in the preceding section if we want to calculate a required sample size. But we should certainly prefer to base a confidence interval on the $\bar{y}$ and s of our data rather than on $\bar{y}$ and an arbitrarily chosen value of σ. What we want is a statement like (5.5.2), but with s replacing σ.

Let us track (5.5.2) back to its beginning. That beginning was the statement (5.4.2):

$$\frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \begin{cases} = Z & \text{if } Y \text{ is normal,} \\ \approx Z & \text{if } Y \text{ is nonnormal and } n \text{ is large.} \end{cases}$$

If we now plan to replace σ by s in our calculations, then there are *two* random variables in the fraction on the left: $\bar{Y}$ and s. And our argument about a confidence interval for μ starts with the fraction

$$\frac{\bar{Y} - \mu}{\frac{s}{\sqrt{n}}}. \tag{5.7.1}$$

This fraction does *not* have the probability distribution of Z.

When n is "large," the fraction behaves in a manner close to that of Z, and for many years statisticians had to rely on that approximation. But results were always dubious when small values of n were involved, and such cases became the frequent ones where the time, expense, or other difficulty of making observations force the investigator to limit the number of observations. Also there are many situations where an investigator can be satisfied with a fairly wide confidence interval, based on few observations for economic reasons, provided he can depend on the level of confidence.

In the early years of this century, William S. Gossett, a mathematician working for the Guinness Brewery in Great Britain, worked out the exact probability distribution of the fraction (5.7.1) in the case where the population of Y has a normal distribution. He published his results under the pseudonym "Student." (There is a widely held belief that he was forced to use a pseudonym because his employers looked on the mathematical results as a potential trade secret.) The distribution that he derived came to be known as *Student's distribution*. The lower-case letter t became standard notation for the random variable, and now the distribution is customarily referred to as *Student's t distribution*, or simply the *t distribution*.

The t distribution is a continuous distribution having a probability density function that graphs as a smooth curve above a t axis extending from $-\infty$ to $+\infty$. The curve is symmetric about $t = 0$ and looks very much like the curve for the standard normal random variable Z except for being more widely spread. Its precise shape depends on the numerical value of n. Thus there is a different t distribution for every different value of n.

After Gossett's breakthrough concerning the fraction (5.7.1), many other applications of t distributions were discovered. Sample size is involved in all of these, but in various ways for various applications. Hence something other than n is preferable as a parameter to identify the different t distributions. The choice that was made is a parameter called *the number of degrees of freedom* (abbreviated d.f.). The concept is the same one which we discussed in Chapter 3 when taking $n - 1$ as the divisor in s^2. Indeed since s^2 is involved in the fraction (5.7.1), it turns out that the fraction has the *t distribution with $n - 1$ d.f.*

The following diagram illustrates the general shape of t distributions and shows the reduction in spread as the number of degrees of freedom increases, with the standard normal distribution being the limit approached as the number of degrees of freedom goes to infinity.
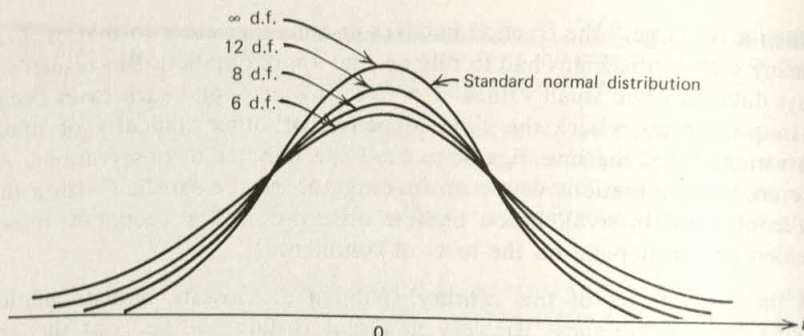
Since there is a different t distribution for every different number of degrees of freedom, it is impractical to give tables for t in the same detail as we have used in Table A-3 for the single distribution of Z. What is done is to table a few of the most useful percentiles of the t distribution for a variety of values of the number of degrees of freedom. Table A-4 in the Appendix is such a table.

The following diagram illustrates the use of Table A-4.
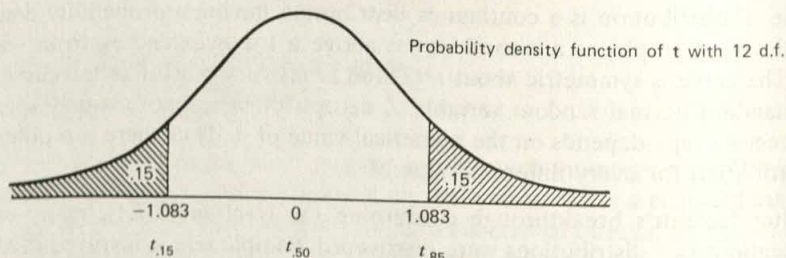
$$P(t \leqslant -1.083 \mid 12 \text{ d.f.}) = .15, \qquad P(t \leqslant 1.083 \mid 12 \text{ d.f.}) = .85,$$
$$P(-1.083 < t < 1.083 \mid 12 \text{ d.f.}) = .70.$$

Applying the probability theory developed by Gossett and others, we arrive at the following operational rule, to replace (5.4.2) when σ is unknown:

$$\frac{\bar{Y}-\mu}{\frac{s}{\sqrt{n}}} \begin{cases} = t \text{ with } n-1 \text{ d.f.} & \text{if } Y \text{ is normal,.} \\ \approx t \text{ with } n-1 \text{ d.f.} & \text{if } Y \text{ is nonnormal and } n \text{ is large} \end{cases} \qquad (5.7.2)$$

From this rule we can proceed in exactly the same manner as in Section 5.5, arriving at a confidence interval to replace (5.5.2).

Given a random sample of n observations from the population of Y, a c percent confidence interval for the population mean μ is

$$\bar{y} - t_* \left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{y} + t_* \left(\frac{s}{\sqrt{n}}\right), \qquad (5.7.3)$$

where $\bar{y}$ = the mean of the sample,
  s = the standard deviation of the sample,
  n = the number of observations in the sample,
  t_* = the confidence-limit value of t with $n-1$ d.f., determined by
     $P(-t_* < t < t_* \mid n - 1 \text{ d.f.}) = c$ percent.

Probability density function of t with $n-1$ d.f.
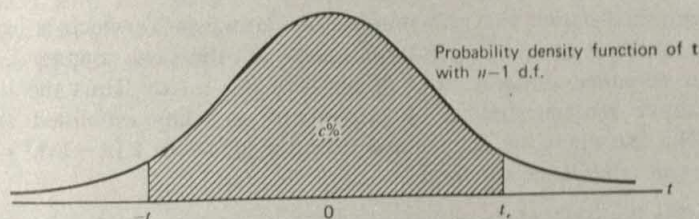
Example 5.7.1

Consider the random sample of nine student G.P.A.'s that we drew in Section 5.3. There we found $\bar{y} = 2.313$, $s^2 = 0.7117$, and $s = 0.844$. Assuming that the G.P.A. is normally distributed, we can apply (5.7.3) to obtain a confidence interval for μ. Let us find a 95 percent confidence interval. Here t has $n - 1 = 9 - 1 = 8$ d.f.

Table A-4 gives us the following probability diagram.

Probability density function of t with 8 d.f.
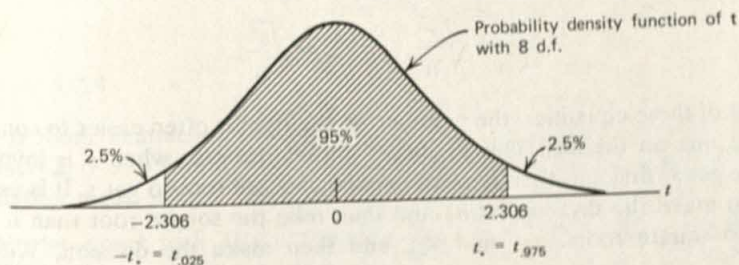
The limits of a 95 percent confidence interval for μ are

$$2.313 \pm 2.306\left(\frac{.844}{\sqrt{9}}\right) = 2.313 \pm 2.306\left(\frac{.844}{3}\right)$$

$$= 2.313 \pm 2.306(.281)$$

$$= 2.313 \pm .648,$$

so that a 95 percent confidence interval for μ is

$$1.665 < \mu < 2.961.$$

If this interval seems unduly wide for practical use, keep in mind that it is based on a sample of only nine observations. If tighter estimation of μ is required, a larger size sample will have to be taken.

We mentioned earlier that the standard deviation of $\bar{Y}$, which is $(\sigma/\sqrt{n})$, is called the standard error of the (sample) mean. In the same manner, $(s/\sqrt{n})$ is called the *estimated standard error of the (sample) mean*. Thus the limits in (5.7.3) can be remembered as "$\bar{y}$ plus–minus so many *estimated* standard errors." The "so many" is determined from the tables of **t** $(n-1$ d.f.) to give the degree of confidence we want.

Comparing the intervals (5.5.2) and (5.7.3), we see that they have exactly the same general structure, with z_* being used when we know σ, and t_* being used when we do not know σ and take s in its place. To keep the distinction straight, it may be helpful to look on (σ, z) as a pair of running mates and on (s, t) as another pair. Having this convenient arrangement of one general structure, easily specialized to two different situations, is one of the rewards of accepting the $(n-1)$ in the definition of s^2. If we had started with n, there would now be unpleasant contortions to bring in the degrees of freedom.

One computational comment should be made. Calculating $(\sigma/\sqrt{n})$ or $(s/\sqrt{n})$ is usually unpleasant since n is not often a perfect square. We should always keep in mind the equalities

$$\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}, \qquad \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}.$$

In either of these equalities, the member on the right is often easier to compute than the one on the left. This is almost always the case when s is involved, since we get s^2 first and then have to extract a square root to get s. It is usually easier to make the division (s^2/n) and then take the square root than it is to take two square roots, $\sqrt{s^2}$ and $\sqrt{n}$, and then make the division. We also

generally get a bonus by having less rounding error. Even in Example 5.7.1, where $n = 9$, we do just as well with $\sqrt{s^2/n}$:

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{.7117}{9}} = \sqrt{.0791} = .281$$

Example 5.7.2

A random sample consisting of 10 rats were placed on a fat-free diet. Their gains in weight were recorded after two weeks, and the mean gain was $\bar{y} = 60$ grams, with a calculated standard deviation $s = 10$ grams. Place 95 percent confidence limits on μ, the true mean gain in weight for the population.

Here t has 9 degrees of freedom, that is, the number of degrees of freedom associated with s^2. For a 95 percent confidence interval, using the t table (Table A-4 in the Appendix) as we did before, we find $t_* = 2.262$, and then we have the 95 percent confidence interval as follows:

$$\bar{y} - t_* \cdot \frac{s}{\sqrt{n}} < \mu < y + t_* \cdot \frac{s}{\sqrt{n}}$$

$$60 - 2.262\left(\frac{10}{\sqrt{10}}\right) < \mu < 60 + 2.262\left(\frac{10}{\sqrt{10}}\right)$$

$$60 - 2.262\sqrt{\frac{100}{10}} < \mu < 60 + 2.262\sqrt{\frac{100}{10}}$$

$$60 - 2.262\sqrt{10} < \mu < 60 + 2.262\sqrt{10}$$

$$60 - 2.262(3.16) < \mu < 60 + 2.262(3.16)$$

$$60 - 7.15 < \mu < 60 + 7.15$$

$$52.85 < \mu < 67.15$$

Therefore the true mean μ of the weight gain on a fat-free diet lies in the interval (52.85 ↔ 67.15 grams). I make this statement with 95 percent confidence.

Example 5.7.3

A random sample of 20 observations from a normal population yields $\bar{y} = 84.7$ and $s^2 = 24.68$. What is a 99 percent confidence interval for the population mean μ?

Here we use **t** with 19 d.f. Taking probability .99 in the center of the t distribution leaves .005 in the tail on the left and .005 in the tail on the right. Hence the t value on the left is $t_{.005}$ (the 0.5th percentile) and the t value on the right is $t_{.995}$ (the 99.5th percentile). Table A-4 shows these values to be -2.861 and $+2.861$.

The limits of a 99 percent confidence interval for μ are

$$84.7 \pm 2.861 \sqrt{\frac{24.68}{20}} = 84.7 \pm 2.861 \sqrt{1.23}$$

$$= 84.7 \pm 2.861(1.11)$$

$$= 84.7 \pm 3.2,$$

giving the 99 percent confidence interval

$$81.5 < \mu < 87.9.$$

Insisting on calculating $(s/\sqrt{n})$ would lead to the following more tedious computation:

$$\frac{s}{\sqrt{n}} = \frac{\sqrt{24.68}}{\sqrt{20}} \approx \frac{\sqrt{24.7}}{\sqrt{20}} = \frac{4.97}{4.47} = 1.11.$$

In Section 5.6 we worked out a procedure for choosing a sample size n that would meet certain specifications for closeness of estimating μ. That procedure requires the use of a known or assumed value of σ. You might now think that we could use (5.7.3) to give us a way of choosing n without assuming a value for σ. But this is impossible. One half of the interval in (5.7.3) is $t_*(s/\sqrt{n})$. If we specify that this should not exceed the amount d, we state

$$\frac{t_* s}{\sqrt{n}} \leq d.$$

But t_* cannot be known without knowing the number of degrees of freedom, and that number $(n-1)$ requires the value of n, for which we are now trying to solve. Moreover s cannot be known until we draw the sample and make the calculations. Thus there is no way to solve the above inequality for n. Hence the best we can do ahead of time is to choose n by the method of Section 5.6. After we draw the sample we should proceed by the method of the present section, getting a confidence interval for μ in accordance with (5.7.3), using t_* and s.

EXERCISES

5.7.1   A sample of 16 from a certain normal population gave the results $\bar{y} = 74.92$ and $s = 12.00$. Give a 95 percent confidence interval for the population mean μ.

5.7.2   A random sample of 10 observations on the temperature of a kiln gave the following results (degrees centigrade): 45, 55, 68, 55, 51, 44, 42, 45, 53, and 37. Determine a 95 percent confidence interval for the true mean kiln temperature.

5.7.3   A sample of 16 from a certain normal population yields $\bar{y} = 124$ and $s = 20$.
a. Give a 99% confidence interval for the population mean μ.
b. What is the precise meaning of "99% confidence interval" in this situation?

5.7.4   A sample of $n = 10$ observations is taken from a large number of accounts receivable. The calculated mean accounts receivable balance of these 10 accounts is $60 with a calculated standard deviation $s = 10. Place 95 percent confidence limits on the true mean accounts-receivable balance.

5.7.5   A random sample of 20 7-ounce "Sticky" shampoo bottles was selected. The net content in each bottle was determined. These are recorded as follows:

$$6.9 \quad 8.0 \quad 7.2 \quad 7.2 \quad 7.4 \quad 7.1 \quad 7.2 \quad 7.0 \quad 6.9 \quad 7.0$$
$$7.0 \quad 7.8 \quad 6.9 \quad 7.1 \quad 7.5 \quad 6.9 \quad 7.4 \quad 7.5 \quad 6.8 \quad 7.2$$

Place 90 percent confidence limits on the true mean net content of a 7-ounce bottle of "Sticky" shampoo.

5.7.6   A large department store had 100,000 accounts receivable at the end of the fiscal year. A random sample of 1700 accounts was taken and the following were calculated:

$\bar{y} = 64$ days (mean age of the accounts)

$s = 25.6$ days (standard deviation of the age of the accounts)

Calculate 95 percent confidence limits on the true mean age of accounts receivable.

5.8 A CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO MEANS, $\mu_1 - \mu_2$, WHEN σ_1 AND σ_2 ARE KNOWN

There are many practical situations where the important point at issue is the *difference* between the means of *two* populations. What is the difference in mean life between two brands of automobile tires? What is the difference in mean reduction of blood pressure accomplished by two different drugs? What is the difference in mean performance of students taught by two different methods? What is the difference in mean lung capacity between smokers and nonsmokers?

In each such case there are two populations involved, one associated with one circumstance, one associated with the other circumstance: tire population 1 for brand 1, tire population 2 for brand 2; population 1 treated with drug 1, population 2 treated with drug 2; student population 1 taught by method 1, student population 2 taught by method 2; population 1 composed of smokers, population 2 composed of nonsmokers. In each case, population 1 has mean and standard deviation μ_1 and σ_1, respectively, while population 2 has mean and standard deviation μ_2 and σ_2, respectively.

Right away there is the serious *nonstatistical* question as to whether the difference $\mu_1 - \mu_2$ makes any practical sense. If tire population 1 is subjected to high-speed daytime driving in the desert while tire population 2 is assigned low-speed nighttime driving in the mountains, the difference $\mu_1 - \mu_2$ in mean life involves a lot more than just a difference in brands. If the population of smokers is composed of old men and the population of nonsmokers is composed of 20-year-old athletes, the difference $\mu_1 - \mu_2$ in mean lung capacity reflects much more than just smoking versus nonsmoking. Statistical inference is no substitute for logical scientific judgment. It is a powerful tool for facilitating decisions in situations where all of the customary rules of sound scientific procedure have been followed.

In scientific experiments, the characteristic observed and measured—the Y in our discussions—is often called the *response* of the population. In the above examples the response is tire life or reduction in blood pressure or student performance or lung capacity. The characteristic by which we identify two or more populations separately—the tire brand, the prescribed drug, the teaching method, the smoking status—is called a *factor*. The different specifications for a factor are called the *levels* of the factor. Thus we can say that we are going to study the response *blood-pressure reduction* under two levels of the factor *drug*, the factor levels being *drug 1* and *drug 2*.

The fundamental rule for a *controlled* experiment is that, except for the response that we shall observe, two populations should differ *only* with respect to the two levels of the factor at issue. This is the notion of having all other factors "controlled." That is why the driving tests for tires will be prescribed

identically for brands 1 and 2; the tires whose lives we will observe should otherwise differ only as to brand. Students taught by method 1 should differ from those taught by method 2 only with respect to method; any other factors that could influence student performance should be operating at the same level in both populations.

Such complete control in experimentation is not always possible. Even when it is possible, careful planning of the experiment may enable us to study the effects of more than a single factor at a time. This is a large scientific subject in itself, and we have no intention of doing anything more here than mention its existence.

This lengthy discussion has been meant to alert the reader to the need for good scientific sense before blindly applying statistical techniques to data. In what follows we *assume* that the two populations under discussion have been defined in this way, and the observations have been made under such conditions that the difference in means $\mu_1-\mu_2$ does make good sense. Our purpose then is to obtain a confidence interval for that difference.

If from population Y_1 (having mean μ_1 and standard deviation σ_1) we construct a random sample of size n_1, then, as we have seen earlier, the sample mean $\bar{Y}_1$ has a probability distribution in which the mean is μ_1 and the standard deviation is $(\sigma_1/\sqrt{n_1})$. Similarly if from population Y_2 (having mean μ_2 and standard deviation σ_2) we construct a random sample of size n_2, the sample mean $\bar{Y}_2$ has a probability distribution in which the mean is μ_2 and the standard deviation is $(\sigma_2/\sqrt{n_2})$. What about the *difference* between the two sample means, $\bar{Y}_1-\bar{Y}_2$?

Since $\bar{Y}_1$ and $\bar{Y}_2$ are random variables, their difference $\bar{Y}_1-\bar{Y}_2$ is a random variable. As such it has a certain probability distribution, with a certain mean and a certain standard deviation. Probability theory has established the following important facts about this distribution.

1. The mean of $\bar{Y}_1-\bar{Y}_2$ is $\mu_1-\mu_2$:

$$\mu_{\bar{Y}_1-\bar{Y}_2}=\mathscr{E}(\bar{Y}_1-\bar{Y}_2)=\mu_1-\mu_2. \tag{5.8.1}$$

2. If the two samples are drawn independently, the standard deviation of $\bar{Y}_1-\bar{Y}_2$ is the square root of the *sum* of the variances of $\bar{Y}_1$ and $\bar{Y}_2$:

$$\sigma_{\bar{Y}_1-\bar{Y}_2}=\sqrt{\mathscr{E}\{[(\bar{Y}_1-\bar{Y}_2)-(\mu_1-\mu_2)]^2\}}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{5.8.2}$$

if $\bar{Y}_1$ and $\bar{Y}_2$ are independent.

As in the case of a single-sample mean, the standard deviation of $\bar{Y}_1-\bar{Y}_2$ is customarily referred to as the *standard error of the difference of (sample) means.*

3. If Y_1 and Y_2 both have normal distributions, then the distribution of $\bar{Y}_1 - \bar{Y}_2$ is normal, for any values of n_1 and n_2; otherwise, the distribution of $\bar{Y}_1 - \bar{Y}_2$ approaches the normal distribution as n_1 and n_2 both go to infinity.

We can now go back to (5.4.2), rephrase it in terms of $\bar{Y}_1 - \bar{Y}_2$, and then carry the discussion forward to a confidence interval in the form of (5.5.2).

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \quad \begin{cases} = Z \text{ if } Y_1 \text{ and } Y_2 \text{ are both normal,} \\ \approx Z \text{ otherwise, when } n_1 \text{ and } n_2 \\ \quad \text{are both large.} \end{cases} \qquad (5.8.3)$$

Given independent random samples of sizes n_1 and n_2, respectively from the populations of Y_1 and Y_2, a c percent confidence interval for the difference $\mu_1 - \mu_2$ between the population means is

$$(\bar{y}_1 - \bar{y}_2) - z_* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{y}_1 - \bar{y}_2) + z_* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \qquad (5.8.4)$$

where symbols have the same meaning as in (5.5.2).

Example 5.8.1

Two normal populations have standard deviations 10 and 15, respectively. Independent random samples were drawn from the populations, 20 observations from the first and 25 from the second. The resulting sample means were 84.7 and 72.1, respectively. What is a 95 percent confidence interval for the difference between the population means?

The standard error of $\bar{Y}_1 - \bar{Y}_2$ is

$$\sqrt{\frac{(10)^2}{20} + \frac{(15)^2}{25}} = \sqrt{\frac{100}{20} + \frac{225}{25}} = \sqrt{5 + 9} = \sqrt{14} = 3.74.$$

The central 95 percent of the distribution of Z is contained between $z = -1.96$ and $z = +1.96$. Thus $z_* = 1.96$. Hence by (5.8.4), the limits of a 95 percent confidence interval for $\mu_1 - \mu_2$ are

$$(84.7 - 72.1) \pm (1.96)(3.74) = 12.6 \pm 7.3,$$

giving the 95 percent confidence interval as

$$5.3 < \mu_1 - \mu_2 < 19.9.$$

Notice that throughout this interval μ_1 is larger than μ_2, since the difference $\mu_1 - \mu_2$ is always positive, going from +5.3 to +19.9. The interval thus allows us to conclude that we are 95 percent confident that μ_1 exceeds μ_2 by an amount somewhere between 5.3 and 19.9.

There can occur intervals with one or two negative boundaries. In such cases we have to draw the appropriate conclusions about the relative sizes of μ_1 and μ_2, as in the following examples.

Example 5.8.2

$$\sigma_1 = 8, \qquad n_1 = 16, \qquad \bar{y}_1 = 40.2;$$
$$\sigma_2 = 6, \qquad n_2 = 9, \qquad \bar{y}_2 = 47.4.$$

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{64}{16} + \frac{36}{9}} = \sqrt{4 + 4} = \sqrt{8} = 2.83.$$

The limits of a 95 percent confidence interval for $\mu_1 - \mu_2$ are

$$(40.2 - 47.4) \pm (1.96)(2.83) = -7.2 \pm 5.5,$$

so that the 95 percent confidence interval is

$$-12.7 < \mu_1 - \mu_2 < -1.7.$$

Here we see that $\mu_1 - \mu_2$ is negative throughout the interval, showing that μ_2 is always larger than μ_1 in the interval. The excess of μ_2 over μ_1 is somewhere between 1.7 and 12.7. It is clearer to express the difference in this direction, a maneuver easily performed by multiplying each member of the inequality by -1:

$$12.7 > -\mu_1 + \mu_2 > 1.7,$$

that is,

$$1.7 < \mu_2 - \mu_1 < 12.7.$$

Example 5.8.3

$$\sigma_1 = 16, \qquad n_1 = 64, \qquad \bar{y}_1 = 120.0$$
$$\sigma_2 = 20, \qquad n_2 = 80, \qquad \bar{y}_2 = 116.0$$

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{256}{64} + \frac{400}{80}} = \sqrt{4 + 5} = \sqrt{9} = 3.$$

The limits of a 95 percent confidence interval for $\mu_1 - \mu_2$ are

$$(120.0 - 116.0) \pm (1.96)\ (3) = 4.0 \pm 5.9,$$

giving the 95 percent confidence interval

$$-1.9 < \mu_1 - \mu_2 < 9.9.$$

In this interval the difference between the means varies from where μ_2 is 1.9 larger than μ_1 to where μ_1 is 9.9 larger than μ_2. The interval includes the case $\mu_1 - \mu_2 = 0$, where μ_1 and μ_2 are equal. We have to end up with the conclusion that we can be 95 percent confident that the relation between the two population means is somewhere between the case where μ_2 is 1.9 larger than μ_1 and the case where μ_1 is 9.9 larger than μ_2. This leaves the case of *no* difference ($\mu_1 - \mu_2 = 0$) within our confidence boundaries.

Example 5.8.4

Let us suppose that there are two different packing lines, number 6 and number 7, producing 1-pound cans of Ohwell Tea. While $\sigma_6 = .03$ ounce is correct for line 6, line 7 is older and has $\sigma_7 = .04$ ounce. During a given day, 16 random cans were taken from each line and the mean packed weights were recorded. The mean for line 6 was 16 ounces, and the mean for line 7 was 16.04 ounces. Place 99 percent limits on the difference between the true line means. The solution is as follows:

Line 7	Line 6
$\bar{y}_7 = 16.04$ ounces	$\bar{y}_6 = 16$ ounces
$n_7 = 16$ cans	$n_6 = 16$ cans
$\sigma_7 = 0.04$ ounce	$\sigma_6 = 0.03$ ounce

Using the procedures previously followed, we find the 99 percent confidence limits for $\mu_7 - \mu_6$ to be

$$(16.04 - 16.00) \pm 2.576 \sqrt{\frac{.0016}{16} + \frac{.0009}{16}} = 0.04 \pm 2.576 \sqrt{.0001 + .00005625}$$

$$= 0.04 \pm 2.576 \sqrt{.00015625}$$

$$= 0.04 \pm (2.576)(.0125)$$

$$= 0.04 \pm .0322,$$

giving

$$0.0078 < \mu_7 - \mu_6 < 0.0722$$

At the 1 percent risk of being wrong, we conclude that there is a real difference between the performance of the two lines, on the average line 7 packing anywhere between .0078 and .0722 ounce more tea per can than line 6.

5.9 A CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO MEANS, $\mu_1 - \mu_2$, WHEN σ_1 AND σ_2 ARE UNKNOWN

As in the situation involving just one population, we generally do not know σ_1 and σ_2. Again large values of n_1 and n_2 can encourage us to replace σ_1^2 and σ_2^2 by s_1^2 and s_2^2, respectively, and take (5.8.4) as an approximation to the confidence interval for $\mu_1 - \mu_2$. But again as in the one-sample case, there are times when important studies have to use small samples. Is there now a procedure employing the t distribution in a manner similar to that of Section 5.7? The answer is *yes*, but there is a new difficulty.

The theory that gives us a t distribution concerning $\bar{Y}_1 - \bar{Y}_2$ not only requires that Y_1 and Y_2 be normally distributed (as in the single-sample case) but also requires that the two populations have the same standard deviation: $\sigma_1 = \sigma_2$. Such equality of standard deviations is a severe restriction, and of course it is not satisfied in general. However, the case where the difference in means $\mu_1 - \mu_2$ makes the most practical sense is the case where the standard deviations are at least approximately equal. Consider the following diagrams of two pairs of distributions.
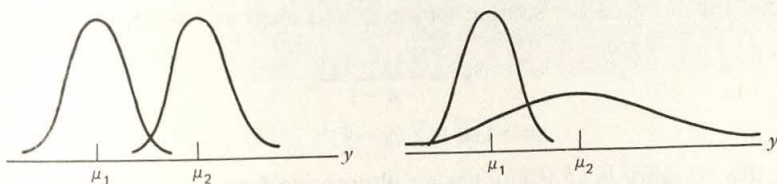
The diagram above shows two distributions having equal dispersions but different means. The difference in means gives a very good measure of the separation of the two chance processes. In the diagram on the right are shown two distributions differing both as to mean and as to dispersion. Here the difference in means is a very incomplete measure of comparison. Notice, for example, that extremely small y values have greater probability in population 2 than in population 1, even though the mean of population 2 is larger than that of population 1.

Because we are concentrating in this book on the most cleancut cases of statistical inference, we shall deal only with the $\sigma_1 = \sigma_2$ situation when comparing two means μ_1 and μ_2 in the absence of precise values for σ_1 and σ_2. If you ask how we can assume that σ_1 and σ_2 are equal when we cannot assume a specific value for either one of them, we must answer that there do exist methods to check the assumption, methods that you would meet in more advanced studies. Such studies would also tell you what to do when the assumption $\sigma_1 = \sigma_2$ cannot be tolerated. Fortunately the standard method that we shall present gives results close to exact when σ_1 and σ_2 differ by a small amount and the sample sizes are equal or nearly so.

Operating under the assumption that $\sigma_1^2 = \sigma_2^2$, we would reasonably want to average s_1^2 and s_2^2 in some way, to produce a single estimate of what is assumed to be the single value of σ_1^2 and σ_2^2 (say σ^2). If the sample sizes n_1 and n_2 are different, we would want to give greater weight to the s^2 that comes from the larger sample. Since we look on the number of degrees of freedom as a measure of the amount of information contained in a sample variance, we shall use degrees of freedom as the weights in our average. We thus produce what is called

a *pooled estimate of variance*, designated s_p^2

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \qquad (5.9.1)$$

As we would deduce from pooling the information in the two samples, s_p^2 has $(n_1-1)+(n_2-1)$ degrees of freedom, that is, n_1+n_2-2 d.f.

Recall the formula for sample variance and clear the fraction:

$$s^2 = \frac{\sum(y-\bar{y})^2}{n-1},$$

$$(n-1)s^2 = \sum(y-\bar{y})^2.$$

Using this equality in (5.9.1) gives an alternative formula which is often more convenient for calculating:

$$s_p^2 = \frac{\sum(y_1-\bar{y}_1)^2 + \sum(y_2-\bar{y}_2)^2}{(n_1-1)+(n_2-1)} = \frac{\sum(y_1-\bar{y}_1)^2 + \sum(y_2-\bar{y}_2)^2}{n_1+n_2-2}. \qquad (5.9.1')$$

This form is easy to remember since it tells us to pool the two sums of squares and pool the degrees of freedom, then make the division.

If we now go back to (5.8.3) and use s_p^2 in place of both σ_1^2 and σ_2^2, we shall have a fraction proved by mathematicians to obey the t distribution with n_1+n_2-2 d.f.:

$$\frac{(\bar{Y}_1-\bar{Y}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{s_p^2}{n_1}+\dfrac{s_p^2}{n_2}}} = t \quad \text{with} \quad n_1+n_2-2 \text{ d.f.}, \qquad (5.9.2)$$

provided Y_1 and Y_2 are normal and $\sigma_1 = \sigma_2$. The equality is a reasonable approximation if violation of the proviso is not excessive and the sample sizes are equal or nearly so.

From (5.9.2) follows a confidence interval to replace (5.8.4) when σ_1^2 and σ_2^2 are unknown.

Given independent random samples of sizes n_1 and n_2, respectively from the populations of Y_1 and Y_2, a c percent confidence interval for the difference $\mu_1-\mu_2$ between the population means is

$$(\bar{y}_1-\bar{y}_2)-t_*\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}<\mu_1-\mu_2<(\bar{y}_1-\bar{y}_2)+t_*\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}, \qquad (5.9.3)$$

where $n_1, n_2 =$ the two sample sizes;
$\bar{y}_1, \bar{y}_2 =$ the two sample means;
$s_p^2 =$ the pooled estimate of variance;
$t_* =$ the confidence-limit value of t with n_1+n_2-2 d.f., determined by $P(-t_*<t<t_*|n_1+n_2-2 \text{ d.f.}) = c$ percent.

The interval is exact or approximate according to the conditions stated with respect to (5.9.2).

The above estimate of standard deviation $\sqrt{(s_p^2/n_1)+(s_p^2/n_2)}$, is commonly called the *estimated standard error of the difference of (sample) means*.

Example 5.9.1
A research investigator reports that he has analyzed independent random samples from two normal populations which have approximately equal variances. His report gives the following information about his data.

$$n_1 = 10, \qquad \bar{y}_1 = 32.7, \qquad s_1 = 4.0;$$

$$n_2 = 12, \qquad \bar{y}_2 = 24.3, \qquad s_2 = 3.5.$$

The pooled estimate of variance, according to (5.9.1), is

$$s_p^2 = \frac{9(4.0)^2+11(3.5)^2}{9+11} = \frac{9(16.00)+11(12.25)}{20}$$

$$= \frac{144.00+134.75}{20} = \frac{278.75}{20} = 13.94.$$

The t distribution here has $10+12-2=20$ (or $9+11=20$) d.f. A 95 percent confidence interval for $\mu_1-\mu_2$ requires $t_*=2.086$ ($t_{.975}$ for 20 d.f. in Table A-4). Applying (5.9.3), we can state that the limits of a 95 percent confidence interval for $\mu_1-\mu_2$ are

$$(32.7-24.3)\pm 2.086\sqrt{\frac{13.94}{10}+\frac{13.94}{12}}=8.4\pm 2.086\sqrt{1.39+1.16}$$

$$=8.4\pm 2.086\sqrt{2.55}$$

$$=8.4\pm 2.086(1.60)$$

$$=8.4\pm 3.3,$$

so that a 95 percent confidence interval for $\mu_1-\mu_2$ is

$$5.1<\mu_1-\mu_2<11.7.$$

When we ourselves handle the sample data from the beginning, we of course use whatever intermediate calculations will simplify the computation of s_p^2. Let us take two very small samples with simple data and carry through the entire procedure for obtaining a 90 percent confidence interval for $\mu_1-\mu_2$.

Example 5.9.2

Sample 1:			Sample 2:	
y_1	y_1^2		y_2	y_2^2
3	9		4	16
1	1		7	49
5	25		9	81
7	49		10	100
4	16		11	121
20	100		4	16
			45	383

$$\bar{y}_1=\frac{20}{5}=4.0 \qquad\qquad \bar{y}_2=\frac{45}{6}=7.5.$$

$$s_1^2=\frac{100-\dfrac{(20)^2}{5}}{4} \qquad\qquad s_2^2=\frac{383-\dfrac{(45)^2}{6}}{5}$$

$$=\frac{100-\dfrac{400}{5}}{4} \qquad\qquad =\frac{383-\dfrac{2025}{6}}{5}$$

$$=\frac{100-80}{4}=\frac{20}{4} \qquad\qquad =\frac{383-337.5}{5}=\frac{45.5}{5}$$

$$=5.00 \qquad\qquad\qquad =9.10.$$

From (5.9.1') we have

$$s_p^2 = \frac{20 + 45.5}{4 + 5} = \frac{65.5}{9} = 7.28.$$

The estimated standard error of $\bar{Y}_1 - \bar{Y}_2$ is

$$\sqrt{\frac{7.28}{5} + \frac{7.28}{6}} = \sqrt{1.46 + 1.21} = \sqrt{2.67} = 1.63.$$

Here t has $5 + 6 - 2 = 9$ (or $4 + 5 = 9$) d.f. For a 90 percent confidence interval t_* is $t_{.95}$, and Table A-4 gives that to be 1.833. Seeing that $\bar{y}_2$ is larger than $\bar{y}_1$, we decide to take the difference in population means as $\mu_2 - \mu_1$, and then have the limits of a 90 percent confidence interval for $\mu_2 - \mu_1$ as

$$(7.5 - 4.0) \pm 1.833(1.63) = 3.5 \pm 3.0,$$

from which it follows that a 90 percent confidence interval for $\mu_2 - \mu_1$ is

$$0.5 < \mu_2 - \mu_1 < 6.5.$$

EXERCISES

5.9.1 Two different automated mechanistic processes were proposed for packaging "Proud" dog food. Let's call them process number 1 and process number 2. Both methods had been extensively tested on a similar product by the manufacturers. Process 1 had $\sigma_1 = 2$ ounces and process 2, $\sigma_2 = 1.8$ ounces. The two processes were installed side-by-side and random samples of "Proud" raw materials were fed into both systems simultaneously. The following packaged weights (in ounces) were obtained from the two processes:

Process 1		Process 2	
58	58	57	57
57	56	58	58
56	57	56	59
58	62	56	58
57	59	56	

a. Determine a 95 percent confidence limit on the difference between the true mean weights produced by the two processes.

b. Is there enough evidence for you to say that the two processes are not doing the same job?

5.9.2 Two vendors have been asked to furnish random samples of "Supercleanser" cans made with a cardboard which will minimize the moisture–vapor transmission rate. The following figures are in grams per 100 square inches per 24 hours.

Supplier A		Supplier B	
.047	.049	.054	.050
.047	.047	.052	.051
.055	.051	.052	.054
.053	.046		

Code the data by multiplying by 1000 and subtracting 40, thus getting 7, 7, 15, 13, and so on.

a. Calculate a 90 percent confidence interval for the true difference between the mean transmission rates.

b. Is there any difference in the cans supplied by the two suppliers?

5.9.3 Twenty new accounting men are to be trained in our accounting system. Two methods of training have been suggested. In order to test these methods, the trainees are divided into two groups. Group 1 is trained by method A and group 2, by method B. After 3 months of training, a test set of material is given to all trainees. The scores of the 20 trainees are shown below.

Training Method A		Training Method B	
y_1		y_2	
96	92	96	94
90	94	84	93
93	83	98	86
88	80	95	89
86	98	91	94

Place a 95 percent confidence interval on the difference between the two training-method means.

5.9.4 Two suppliers are trying to furnish a manufacturer with alkyl–benzene. Each of the suppliers claims that his alkyl–benzene will cause a chemical reaction to yield the highest percentage of completeness of reaction. The percentage of completeness standard deviation for both materials is assumed to be $\sigma = 1$ percent.

The manufacturer did $n_A = 10$ reactions with alkyl–benzene from supplier A and $n_B = 10$ reactions with alkyl–benzene from supplier B. The following data were obtained:

Supplier A		Supplier B	
96.2	95.8	96.3	97.2
95.8	96.4	97.4	94.8
96.3	96.2	95.4	96.8
94.7	97.3	96.7	97.2
95.6	96.8	95.9	97.0

a. Calculate 95 percent confidence limits on the difference between the two percent completeness means.
b. From which supplier would you recommend the manufacturer to buy alkyl-benzene? Why?

5.9.5 a. Using the following sample data, find a 99 percent confidence interval for the population mean.

$$y: \quad 13, 4, 12, 8, 7, 14, 15, 13, 6$$

b. Consider the following data as a sample taken from a second population. Find a 99 percent confidence interval for the population mean.

$$y: \quad 10, 9, 10, 6, 11, 13, 9, 2, 8.$$

c. Give an estimate of the common variance of the two populations considered above.
d. Determine a 99 percent confidence interval for the difference between the two population means.

5.9.6 All accounting information is available from two plants. From plant A, 10 random accounts were selected and checked for percent of errors. From plant B, eight accounts were chosen at random. The results are as follows:

Plant A		Plant B	
12.4	12.1	13.6	12.5
10.6	12.0	12.4	13.6
11.8	12.1	11.8	12.8
12.4	11.8	11.9	13.2
11.9	11.5		

a. Would you consider these plants different? Why?
b. Why do you suppose only eight were taken from plant B?

5.9.7 An analytic study reports the following data on measurement of a certain characteristic in treated and untreated water from a given basic supply, 10 water

	Treated	Untreated
Sample size	10	10
Sample mean	92.46	98.16
Sample variance	36.00	45.00

samples having been used in each case. The characteristic under study is one which is generally regarded as being normally distributed. From the context of the report it appears that the two sets of data were independent random samples. Using 95 percent confidence limits, determine whether the true "treated" mean is significantly smaller than the "untreated" mean. (Show clearly all steps in your decision procedure.)

5.10 A CONFIDENCE INTERVAL FOR THE BINOMIAL PROPORTION p

In Chapter 4 we considered the chance process in which there are only two possible outcomes, such as "success" and "failure." The probability distribution governing this process is the *binomial* distribution, and the random variable is the number of successes in n independent trials of an experiment wherein the probability of success in any single trial is p.

This distribution applies to a wide variety of practical stituations. Whenever the characteristic we want to study in a population is of the two-category kind, we have to deal with the binomial distribution. Such characteristics are very frequently the objects of important investigations: (a) voter opinion in a referendum (for, against), (b) status of a manufactured item produced under quality control (acceptable, not acceptable), (c) condition of a cancer patient 5 years after diagnosis (alive, dead), (d) sex of high-school teacher (male, female).

For any such characteristic, the experiment of drawing at random a member of a population and observing to which of the two categories he (she, or it) belongs is conceptually the same as tossing a coin and observing head or tail in the result. *Head* can designate one of the two categories of the characteristic and *tail*, the other category. In general the coin is biased, having p as the probability of tossing *head* (the "success" of our earlier discussion).

In any such situation our interest centers on the value of p, the probability that a single trial will give *head*. In most practical applications, we want to interpret this probability in its long-run sense—the *proportion* of an unending sequence of tosses that will be *heads*. If a finite population is then large enough, we can look on p as the proportion of the population that belongs to

the category *head*: (a) the proportion of voters who are in favor of the referendum proposition, (b) the proportion of acceptable items coming off the production line (industrial quality control often looks at this from the opposite side, "percent defective"), (c) the 5-year survival rate of cancer patients, (d) the percentage of male high-school teachers.

When we have a *sample* of *n* members drawn from a population, we check out each member of the sample, determining which of the two categories applies, tally the number of members in the "head" category, divide this number by *n*, and thus arrive at the *sample proportion* for "head." This sample proportion we shall designate $\hat{p}$. (There are various ways of speaking the symbol $\hat{p}$: "*p* circumflex" would be precise but is never heard; "*p* hat" is the common expression in this country, with "*p* roof" being the favorite in certain other regions.)

Statistical estimation of *p* proceeds in the same manner as estimation of a mean. It is required that the sample be a *random sample* in the same sense we discussed earlier. It is only to the $\hat{p}$ of such a sample that we can apply the methods of inference which we shall discuss. Hence we define $\hat{p}$ in terms of a random sample.

> Given a random sample of *n* observations from a population, the sample *proportion* $\hat{p}$ is the observed fraction
>
> $$\hat{p} = \frac{\text{number of sample elements in category "head"}}{n}. \qquad (5.10.1)$$

Example 5.10.1

A random sample of 200 women from a large population of women operated on for removal of breast cancer was observed over a period of time covering 5 years after each woman's operation. During the period none of the women died of any cause other than cancer. Ten of the women died of cancer within 5 years after the operation. The observed *sample 5-year survival rate for surgically treated breast cancer* is then

$$\hat{p} = \frac{190}{200} = 0.95.$$

This proportion, as in the case of any such fraction, can be expressed as 95 percent, or 95 per hundred, or 950 per 1000, or in any other form adopted as conventional usage in a specific field of study.

Since a random sample is a chance process, the *number of sample elements in category "head"* is a random variable. Hence so is the *sample proportion* $\hat{p}$. When we draw one random sample, we get one observation on this random variable $\hat{p}$—the *observed* sample proportion $\hat{p}$. In order to make statistical inferences from such an observed value, we need to know the probability behavior of $\hat{p}$ from sample to sample—the probability distribution of $\hat{p}$.

Dealing with the *exact* probability distribution of $\hat{p}$ is beyond the scope of this book. Fortunately there is an approximate distribution that will give us satisfactory results in most practical situations. We can arrive at this by the theory that we have already applied to the sample mean $\bar{Y}$.

The sample proportion $\hat{p}$ *is* a sample mean. In our two-category situation, the population random variable Y is the binomial random variable with parameter p and $n = 1$. Remember that a population random variable is the random variable that applies to any single member chosen at random from the population. So in the present case we have Y as the binomial random variable for a single trial. Y is thus the number of "successes" (or "heads") in one trial. Obviously the only possible values for Y are 1 and 0; in a single trial we can only get one head or no head. Now when we have a random sample of size n, we have the n random variables $Y_1, Y_2, \ldots, Y_n$. Our observations give us the collection $y_1, y_2, \ldots, y_n$, wherein each y_i is either 1 or 0—1 if outcome is "head," 0 if outcome is "tail." When we add the y values, the total is precisely the number of "heads" in the sample. Hence the sample mean $\bar{Y}$ is

$$\bar{Y} = \frac{\sum Y}{n} = \frac{\text{number of sample elements in category "head"}}{n} = \hat{p}.$$

Making this interpretation of $\hat{p}$ as a sample mean $\bar{Y}$, we can use all of the facts in Section 5.4.

First we refer to (4.8.5) and find that, since our present Y is binomial with $n = 1$,

$$\mu = (1)p = p,$$
$$\sigma = \sqrt{(1)pq} = \sqrt{pq}, \quad \text{where } q = 1 - p.$$

Using these values in (5.4.1), we obtain

$$\mu_{\bar{Y}} = \mu = p,$$
$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{pq}}{\sqrt{n}} = \sqrt{\frac{pq}{n}}.$$

But the $\bar{Y}$ here is our sample proportion $\hat{p}$. Thus we have the following important facts about $\hat{p}$.

If $\hat{p}$ is the proportion of "heads" in a random sample of size n drawn from a population in which the probability of a member's being "head" is p, then the probability distribution of $\hat{p}$ has the following mean and standard deviation:

$$\begin{cases} \text{Mean of } \hat{p}: & \mu_{\hat{p}} = p, \\ \\ \text{Standard deviation of } \hat{p}: & \sigma_{\hat{p}} = \sqrt{\dfrac{pq}{n}}, \text{ where } q = 1 - p. \end{cases} \qquad (5.10.2)$$

Next we can apply the central limit theorem (Theorem 5.4.2) since $\hat{p}$ is a $\bar{Y}$. That gives us the information we need for making an approximation to the probability distribution of $\hat{p}$. If $\hat{p}$ is the sample proportion referred to in (5.10.2), then as n increases without bound, the distribution of $\hat{p}$ approaches the normal distribution with mean p and standard deviation $\sqrt{pq/n}$.

How fast the distribution of $\hat{p}$ approaches the normal distribution as n increases depends on the value of p. The approach is fastest when $p = \frac{1}{2}$; it is very slow when p is near 0 or 1. It has been found that the normal distribution is a satisfactory approximation to the distribution of $\hat{p}$ if $np > 5$ and $nq > 5$.

The above discussion gives us the operational rule:

$$\frac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}} \approx Z \quad \text{if } np > 5 \text{ and } nq > 5, \text{ where } q = 1 - p. \qquad (5.10.3)$$

Just as the operational rule (5.4.2) led to the confidence interval (5.5.2) for μ, the same logic will now give us the confidence interval

$$\hat{p} - z_* \sqrt{\frac{pq}{n}} < p < \hat{p} + z_* \sqrt{\frac{pq}{n}}. \qquad (5.10.4)$$

But now a new difficulty has appeared: the unknown p is not only in the interior of the interval, where we want it, but also in the limits, where we definitely do *not* want it. The limits must be expressible completely in terms of z_* and *observed* data or else we cannot calculate them.

We can apply some algebraic manipulations to the above interval and end up with interval limits involving only z_*, $\hat{p}$, and n. Those limits will be fairly complicated expressions, and there is a question whether the approximation we are using justifies the refined effort. The procedure that is customarily followed is to settle for an additional approximation by using $\hat{p}$ in place of p in the standard deviation $\sqrt{pq/n}$. We thus decide on the following statement of a confidence interval for p.

Given a random sample of n observations on a population in which the probability of a member's being "head" is p, an approximate c percent confidence interval for p is

$$\hat{p} - z* \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z* \sqrt{\frac{\hat{p}\hat{q}}{n}}, \qquad (5.10.5)$$

where n = the number of observations in the sample;
$\hat{p}$ = the proportion of "heads" in the sample, $\hat{q} = 1 - \hat{p}$;
$z*$ = the confidence-limit value of z, determined by

$$P(-z* < Z < z*) = c \text{ percent},$$

and the approximation is satisfactory if $n\hat{p}$ and $n\hat{q}$ are both greater than 5.

As in other cases of sample means, the standard deviation $\sqrt{pq/n}$ is usually referred to as the *standard error of the (sample) proportion*, and the estimated standard deviation $\sqrt{\hat{p}\hat{q}/n}$ is called the *estimated standard error of the (sample) proportion*.

Example 5.10.2

What is (approximately) a 99 percent confidence interval for the 5-year survival rate in the population for which Example 5.10.1 gave a sample estimate?

The data in Example 5.10.1 gave $n = 200$ and $\hat{p} = .95$. We note that $n\hat{p} = 200(.95) = 190$, $n\hat{q} = 200(.05) = 10$; since both of these values are greater than 5, we judge (5.10.5) to be an acceptable approximation. For 99 percent confidence, $z* = 2.58$ ($z_{.995}$ in Table A-3). Thus the limits of an approximate 99 percent confidence interval for p are

$$.95 \pm 2.58 \sqrt{\frac{(.95)(.05)}{200}} = .95 \pm 2.58 \sqrt{\frac{.0475}{200}}$$

$$= .95 \pm 2.58 \sqrt{.000238}$$
$$= .95 \pm 2.58(.0154)$$
$$= .95 \pm .0397,$$

so that an approximate 99 percent confidence interval for p is

$$.9103 < p < .9897,$$

or, in percent notation,

$$91.0 \text{ percent} < p < 99.0 \text{ percent}.$$

It is of interest to report that the more advanced methods which use the exact distribution of $\hat{p}$ would here give us

$$89.7 \text{ percent} < p < 98.2 \text{ percent.}$$

5.11 REQUIRED SAMPLE SIZE FOR ESTIMATING A PROPORTION p

In Section 5.6 we considered how to choose a sample size so that we could have a specified degree of confidence that the resulting sample mean $\bar{y}$ will be within a specified distance d of the population mean μ. The same question of what sample size to use comes up in the case where we are estimating a population proportion p. The same argument applies; we want to specify that the come-and-go in a confidence interval shall not exceed d.

The come-and-go in the interval (5.10.5) is of no use ahead of time since it involves $\hat{p}$, which we cannot know until after observing the sample. The come-and-go in the interval (5.10.4) gives us a logical rule:

$$z*\sqrt{\frac{pq}{n}} \leq d, \tag{5.11.1}$$

but this involves p, the very value that we are trying to estimate. If we really could assume a value for p, there would not be any reason to bother with estimation at all.

We do have one mathematical fact to help us. Because both p and q are between 0 and 1, there is a limit to the value that pq can have under any circumstances. If we use this maximum value in (5.11.1) we can be assured that the resulting value of n is large enough regardless of the true value of p. In actual situations we may be able to do somewhat better. Our procedure goes according to the following argument.

The graph of the quantity pq as a function of p is shown in the diagram: (overleaf). Here we see that pq never exceeds 1/4, and that that value occurs when $p = 1/2$. Hence if we use $pq = 1/4$ in (5.11.1) we are bound to get a value of n large enough to suit our requirement. [Recall the effect of the standard deviation in (5.6.1').] As a matter of fact, the value of n will almost always be larger than necessary, for the graph shows that the value of pq decreases away from 1/4 as soon as p moves away from 1/2.
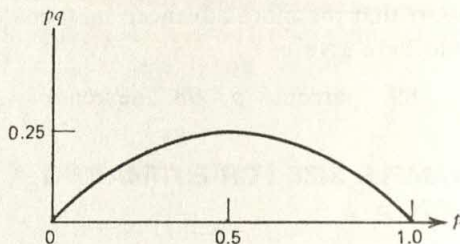
In most practical situations we have some idea of at least a bracket for p. A reasonable procedure then is to takè this bracket for p on the p axis in the above diagram and use the largest value of pq that occurs above the bracketed p interval. The following are some representative examples.

a. $0 < p < .4$: pq rises steadily from 0 over this p interval, reaching its maximum when $p = .4$; that maximum is $pq = (.4)(.6) = .24$.

b. $.4 < p < .7$: pq both rises and falls over this interval; its maximum is the overall maximum for pq: $1/4$.

c. $.7 < p < 1$: pq is decreasing all the way from where $p = .7$ to where $p = 1$, so that its maximum occurs where $p = .7$, that maximum being $pq = (.7)(.3) = .21$.

Example 5.11.1

An investigator interested in the 5-year survival rate of women who had had breast-cancer operations is unsatisfied with the width of the 99 percent confidence interval in Example 5.10.2 ($.91 < p < .99$). He would like to estimate p to within .02 with 99 percent confidence. His experience suggests that p is at least .9. How large a sample is required to meet his estimation specifications?

We see that the maximum value of pq over the interval $.9 \leqslant p < 1$ is the value at $p = .9$: $pq = (.9)(.1) = .09$. Using this value in (5.11.1), we have

$$2.58\sqrt{\frac{(.9)(.1)}{n}} \leqslant .02,$$

$$\frac{2.58\sqrt{.09}}{\sqrt{n}} \leqslant .02,$$

$$2.58\sqrt{.09} \leqslant .02\sqrt{n},$$

$$129\sqrt{.09} \leqslant \sqrt{n},$$

$$16641(.09) \leqslant n,$$

$$1497.69 \leqslant n,$$

giving 1498 as the minimum required sample size.

5.12 A CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS, $p_1 - p_2$

As in the case of the quantitative characteristics we studied in Sections 5.8 and 5.9, there is often practical interest in the difference between two proportions: (a) difference between men and women as to proportion favoring a referendum proposition, (b) difference in percent defective between the outputs of two different machines, (c) difference in 5-year survival rates of patients given two different kinds of operation, (d) difference in percentages of men on high school faculties in public and private schools.

In Section 5.10 we arrived at the probability distribution of a sample proportion $\hat{p}$ by looking on $\hat{p}$ as a sample mean $\bar{Y}$. So now, with independent samples from two populations, we can treat $\hat{p}_1-\hat{p}_2$ as a difference between sample means, and then apply what we had in Section 5.8. The results are as follows, matching the facts (1), (2), and (3) in Section 5.8.

1. The mean of $\hat{p}_1-\hat{p}_2$ is p_1-p_2:

$$\mu_{\hat{p}_1-\hat{p}_2}= p_1 - p_2. \tag{5.12.1}$$

2. If the two samples are drawn independently, the standard deviation (standard error) of $\hat{p}_1 - \hat{p}_2$ is the square root of the sum of the variances of $\hat{p}_1$ and $\hat{p}_2$, these variances being given by (5.10.2):

$$\sigma_{\hat{p}_1-\hat{p}_2}= \sqrt{\frac{p_1 q_1}{n_1}+\frac{p_2 q_2}{n_2}} \tag{5.12.2}$$

if $\hat{p}_1$ and $\hat{p}_2$ *are independent.*

3. The distribution of $\hat{p}_1 - \hat{p}_2$ approaches the normal distribution as n_1 and n_2 both go to infinity.

As in the case of a single proportion, the normal approximation suggested in (3) is generally satisfactory if $n_1 p_1$, $n_1 q_1$, $n_2 p_2$, $n_2 q_2$ are all greater than 5. We are then led to the following version of (5.8.3) applied to sample proportions:

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{n_1}+\dfrac{p_2 q_2}{n_2}}} \approx Z \qquad \begin{array}{l}\text{if } n_1 p_1,\ n_1 q_1,\ n_2 p_2,\ n_2 q_2 \\ \text{are all greater than 5.}\end{array} \tag{5.12.3}$$

There is no theory like that of the t distribution to give us a way of amending (5.12.3) as we amended (5.8.3) when we worked in Section 5.9. Hence in order to obtain a confidence interval for $p_1 - p_2$, we are forced to proceed as we did in Section 5.10, to admit an additional approximation by using $\hat{p}_1$ and $\hat{p}_2$ for p_1 and p_2 in the denominator of (5.12.3). We thus decide on the following statement of a confidence interval for $p_1 - p_2$:

Given independent random samples from two populations having binomial proportions p_1 and p_2, respectively, the respective sample sizes being n_1 and n_2, an approximate c percent confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) - z_* \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_* \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}},$$

(5.12.4)

where $\hat{p}_1 =$ the proportion of "heads" in sample 1,
$\hat{p}_2 =$ the proportion of "heads" in sample 2,
$\hat{q}_1 = 1 - \hat{p}_1,$ $\hat{q}_2 = 1 - \hat{p}_2,$
$z_* =$ the confidence-limit value of z, determined by

$$P(-z_* < Z < z_*) = c \text{ percent},$$

and the approximation is satisfactory if $n_1 \hat{p}_1$, $n_1 \hat{q}_1$, $n_2 \hat{p}_2$, $n_2 \hat{q}_2$ are all greater than 5.

Example 5.12.1

The marketing organization for product T asked each of 70 housewives in a random sample from city 1 whether she had ever tried product T; 49 replied "yes." In a random sample of 120 housewives in city 2, 48 "yes" replies were given to the same question. What is an approximate 95 percent confidence interval for the difference between the two cities with respect to the percentage of housewives who have tried product T?

We let p be the proportion of housewives in a city who have tried product T. Then we have

$$\hat{p}_1 = \frac{49}{70} = .7; \qquad \hat{p}_2 = \frac{48}{120} = .4.$$

The estimated standard error of the difference $\hat{p}_1 - \hat{p}_2$ is

$$\sqrt{\frac{(.7)(.3)}{70} + \frac{(.4)(.6)}{120}} = \sqrt{\frac{.21}{70} + \frac{.24}{120}} = \sqrt{.003 + .002} = \sqrt{.005} = .0707.$$

The confidence-limit value z_* for 95 percent confidence is 1.96, since

$$P(-1.96<Z<1.96)=.95,$$

and so the limits of an approximate 95 percent confidence interval for p_1-p_2 are

$$(.7-.4)\pm1.96(.0707)=.3\pm.139,$$

from which the approximate 95 percent confidence interval for p_1-p_2 is

$$.161<p_1-p_2<.439.$$

Expressed in terms of percentage, the interval is

$$16.1 \text{ percent}<p_1-p_2<43.9 \text{ percent},$$

and we could report that the sampling indicates a higher percentage of housewives in city 1 have tried product T, and, with about 95 percent confidence, we think that the difference in percentage between the two cities is somewhere between 16 and 44.

EXERCISES

5.12.1 A weights-and-measures inspector for consumer products went into a local grocery store to check on the weight of synthetic detergents. She walked to the shelf containing "Ebb" detergent. Each carton was cited as containing 20 ounces of Ebb. The inspector chose $n = 60$ cartons, emptied the contents of each carton onto scales, and read the weight. She found six cartons whose contents were under 20 ounces; the rest were over 20 ounces.
   a. What is the estimate of the true percent under the marked weight of 20 ounces?
   b. Place 95 percent confidence limits on the true percent under marked weight.
   c. Discuss the implications of using this sample as indicative of all Ebb 20-ounce cartons.
   d. What additional information available would help in assessing the percentage of a year's supply under weight?

5.12.2 In a random sample of 100 of a certain kind of seed there were 20 seeds that germinated. Give a 95 percent confidence interval for the number of seeds that will germinate if 400 are planted.

5.12.3 A certain treatment is found effective in 16 out of 25 cases. Construct an approximate 99 percent confidence interval for the probability that the treatment is effective in a single case.

5.12.4 In a random sample of 100 articles produced in a certain process, 10 are found to have defects. Construct a 95 percent confidence interval for the defect rate of the process.

5.12.5 To estimate the proportion of a population in favor of a certain proposal, a statistician questioned a random sample of 200 and found 80 favorable replies. Give a 99 percent confidence interval for the true population proportion p in favor of the proposal.

5.12.6 In a certain achievement test 45 students out of 600 in district A and 39 out of 800 in district B received scores in the "superior" category. Determine a 95 percent confidence interval for the difference between "superior" proportion in the two districts.

5.12.7 In Trenton, New Jersey 148 men and 152 women were polled on the question "Do you approve, by and large, of the practice of tipping?" Eighty-nine men and 116 women replied "yes." Determine a 90 percent confidence interval for the difference between the true approval rates of the two sexes. What inferences can you draw from this?

5.12.8 A local grocery chain received a truck load of 5000 filled peanut-butter jars. In unloading the truck, several of the jars were found to have loose lids. If this were true of most of the jars, there would be a serious problem of spoiled peanut butter. Thus the manager decided to take a random sample of 200 jars and test them for loose lids. The test showed that six of the jars had loose lids.
a. What is your best estimate of the percentage of loose lids in the shipment?
b. Place 95 percent confidence limits on the true percentage of loose lids.
c. Discuss how one would select 200 jars of peanut butter from the truckload of jars. Do you foresee any practical difficulties in doing this?

5.12.9 What sample size will be surely large enough so that we can be 99 percent confident that the resulting $\hat{p}$ will not differ from the true proportion p by more than .025? How does your answer change if you can assume that p is certainly no larger than 0.1?

5.12.10 A shipment of 1200 flashlights was received from Hong Kong. The flashlights were to be used as prizes in a contest. The contest manager decided he'd better try a few to see if they would light. He chose a sample of $n = 100$ flashlights and tested them.
a. How would you choose the 100 flashlights? Illustrate the procedure by selecting the first 10 flashlights.
b. Assume that seven out of the 100 would not light. What is your estimate of the number of defective flashlights in the shipment of 1200?
c. Place 95 percent confidence limits on the true percent defective.
d. What are the 95 percent confidence limits on the number of defective flashlights in the shipment?

5.12.11 A large accounting firm handled the billing for several small companies. Last year a total of 8000 bills were processed by the firm for the Handy Company. After receiving several complaints about billing errors, the Handy Company decided to take an audit on the bills (invoices). The auditor took a random sample of 400 invoices and found eight billing errors.
a. What is your estimate of the billing-error rate?
b. Place 90 percent confidence limits on the true billing-error rate.
c. How many bills would you estimate to have errors?
d. What would you tell the accounting firm if you were president of the Handy Company?

6.1 THE ROLE OF STATISTICS IN THE SCIENTIFIC METHOD

The procedure of investigation that has come to be known as "the scientific method" is a procedure of constructing a hypothesis about a condition or process or law of nature and then checking it against reality by means of observations. One forms his hypothesis, makes observations related to it in the real world, compares the results with what the hypothesis hypothesizes, and then accepts or rejects the hypothesis according as to whether the observed results match the hypothesis.

One of the first dramatic examples we learned about in childhood was Columbus's test of the hypothesis that the earth is a sphere ("the world is round"). By sailing west from Spain he would, according to the hypothesis, eventually reach the Far East. Confirming or denying the hypothesis was delayed for a time since he bumped into the Americas on the way. The movement of the earth around the sun, rather than vice-versa, was for man at first a hypothesis; so was the pattern of circulation of the blood in the human body, the effectiveness of vaccination, the action of penicillin, and Einstein's famous equation $E = mc^2$.

When a hypothesis about a process in Nature has to do with one of Nature's *chance* processes, it is generally not a straightforward matter to check the hypothesis against the observations. In any

To Reject Or Not To Reject
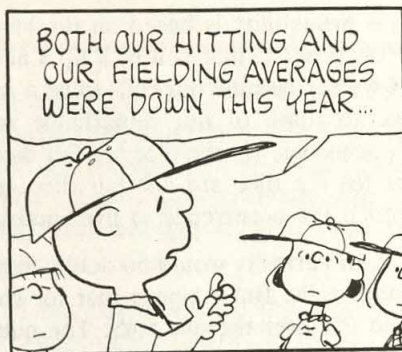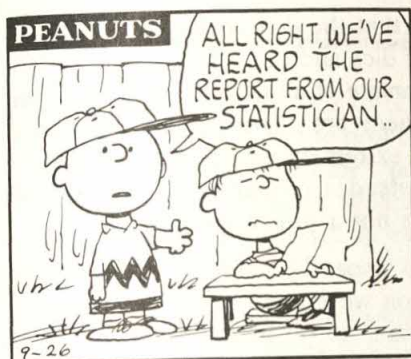
chance process the various possible observations will occur according to the probability distribution of the process. There is natural variability among such observations, the different outcomes happening in the various proportions specified by the probability law. Thus when we use a sample of observations to test a hypothesis, we have to face the fact that the observations will not fit the hypothesis perfectly even if the hypothesis is true.

Looking at the discrepancy between hypothesis and observations, the investigator has to sort out the difference that can be attributed to chance and the difference that must be charged to the falsity of the hypothesis. As in other studies involving chance processes, we can never know the *truth*; all we can do is make decisions according to procedures which have *good odds* of producing correct decisions. It is in such decision-making that statistical inference has a role to play.

Recall the example in Chapter 4 where we calculated the probability of rolling "7" five times in succession with a pair of fair dice. That probability is $(1/6)^5 = .0001286$, indicating that in the long run such an all-7 quintuple of rolls will occur about 13 times out of 100,000, on the average. As specified,

this probability is based on the assumption that the dice and rolls are "fair." Now suppose that you pick up a new pair of dice and at once roll 7, 7, 7, 7, 7. If we can assume our rolls to be a random sample of all possible rolls, what are we to think of our remarkable result? There are logically two alternative conclusions: (a) the dice are fair dice and an extremely rare event has occurred or (b) the dice are not fair dice, having instead a probability distribution in which the occurrence of five successive 7s is not a rare event.

Our curiosity would no doubt compel us to suspend judgment and make a lot more rolls. But suppose that for some reason we *had* to choose between (a) and (b) after the five rolls. The natural reaction to (a) is that commonly used expression, "It's possible but not probable." It does not seem reasonable that the outcome of a single sample from a population should be one of the outcomes that had a very small chance of occurring. It *can* happen but it is very unlikely. With such a point of view, we would choose (b) as our conclusion; that is, we would *reject the hypothesis* that the dice are fair. We would grant that there is a risk of our being wrong, but we would judge the risk to be very small.

This is the rationale of the *statistical test of hypothesis* or *test of statistical significance*. In practice, the word "statistical" is generally omitted from both of these expressions, but its presence should always be clearly understood, because the test is a procedure of *statistical* inference, with conclusions phrased in terms of probability.

6.2 THE LEVEL OF SIGNIFICANCE

In order to *control* the risk of rejecting the hypothesis when it is in fact true, we choose ahead of time the probability boundary below which we are going to call an experimental outcome "possible but not probable." Commonly used values are 5 and 1 percent. Such a probability value, chosen in advance, gives us our definition of how rare an experimental result will have to be in order to make us consider it rare beyond the credibility of chance.

In order to fix ideas by an example free of mathematically technical difficulties, consider the following situation. In a certain manufacturing process, the production line includes a weighing machine set to deliver 12 ounces of the manufactured material into a cardboard box. Boxes thus filled are produced in large quantity and marketed with a label declaring "contents: 12 ounces." Long experience has shown that the weighing machine does not deliver precisely 12 ounces to each and every box, but that the amount delivered is a random variable that is normally distributed, with standard deviation small and very stable over time ($\sigma = 0.1$) but the mean (μ) subject to shifts away from the preset 12 ounces. The quality control section of the company makes a periodic check by taking a random sample of 25 filled boxes and weighing the contents on an elaborately calibrated scales.

The question is: How far below 12 must the mean $\bar{y}$ of the sample of 25 boxes be in order to justify a decision that the mean μ of the production line weighing machine is less than 12? The company hypothesizes that $\mu = 12$, against the alternative hypothesis that $\mu < 12$. It has become customary to refer to the basic hypothesis as the *null hypothesis* and to use for it the notation H_0, suggesting the idea of "no difference," in our case no difference between the process mean μ and the specification value 12 (ounces). We can designate the alternative hypothesis by H_A, and thus label our test as follows:

$$H_0: \ \mu = 12 \quad \text{versus} \quad H_A: \ \mu < 12.$$

The company wants to run only a small risk of rejecting H_0 (and thus accepting H_A) if H_0 is in fact true. For the sake of example we suppose the limit of tolerable risk to be 5 percent. We then argue that the $\bar{y}$ values that justify rejecting H_0 are those far enough below 12 as to have no more than a 5 percent chance of occurring if μ is in fact 12. Any such $\bar{y}$ value will be considered as "significantly" less than 12, the term *significant* being construed in a technical statistical sense as meaning "beyond reasonable attribution to chance." It is in relation to acting as the criterion for this judgment that the 5 percent is called the *level of significance*. A common general notation is the lower-case Greek letter "alpha," so that we speak of $\alpha = .05$ as being the same as the 5 percent level of significance.

6.3 THE CRITICAL REGION

The $\bar{y}$ values far enough below 12 to cause us to reject the null hypothesis that μ is 12 make up a set of values called the *critical region* of the test. The values are "critical" in the sense that if the sample produces any one of them it causes H_0 to be rejected. The probability α associated with the set of critical values is referred to as the *size* of the critical region; it is "size" in the sense of probability mass.

According to Theorem 5.4.1, the $\bar{y}$s of samples of size 25 in our present situation will be distributed in a normal probability pattern having the population mean μ as its mean and the standard deviation $\sigma/\sqrt{n}$, specifically $(0.1/\sqrt{25}) = (0.1/5) = 0.02$. That is, $\bar{Y}$ is normal with mean μ and standard deviation 0.02. Now *if H_0 is true*, then $\mu = 12$ and hence $\bar{Y}$ is normal with mean 12 and standard deviation 0.02.

We can then determine the critical region specifically, as follows.
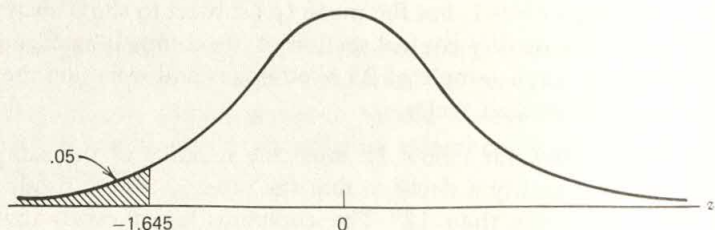
$$\alpha = 0.05 = P(Z \leqslant -1.645)$$
$$= P\left(\frac{\bar{Y}-12}{0.02} \leqslant -1.645\right)$$
$$= P(\bar{Y}-12 \leqslant -0.0329) = P(\bar{Y} \leqslant 11.9671).$$

Thus *if H_0 is true*, there is just 5 percent probability that the $\bar{Y}$ of a sample will turn out to be 11.97 or less. Hence the critical region of the test is the interval $\bar{y} \leqslant 11.97$; if $\bar{Y}$ falls in this region we shall reject H_0.

We can specify the above critical region either directly in terms of $\bar{y}$ or indirectly in terms of z:

$$\bar{y} \leqslant 11.97$$

or

$$z \leqslant -1.645, \qquad \text{where} \qquad z = \frac{\bar{y}-12}{0.02}.$$

The critical region is customarily stated in the manner of a *decision rule*: reject H_0 if $\bar{y} \leqslant 11.97$; accept H_0 otherwise. This rule could be stated also in terms of z:

If H_0 is true, then

$$\frac{\bar{Y}-12}{0.02} = Z;$$

reject H_0 if $z \leqslant -1.645$; accept H_0 otherwise.

General Procedure	*Example*
STEP 1. IDENTIFICATION OF TEST	
a. H_0 versus H_A	H_0: $\mu = 12$ versus H_A: $\mu < 12$.
b. Assumptions	Y is $N(\mu, 0.1)$.
c. Nature of sample	Random sample, $n = 25$.
d. Level of significance	$\alpha = .05$.
STEP 2. THE TEST STATISTIC	
a. Identification of statistic	Test is based on sample mean $\bar{Y}$.
b. Distribution under H_0	If H_0 is true, then $$\frac{\bar{Y}-12}{\sqrt{\frac{.01}{25}}} = Z.$$
STEP 3. THE DECISION RULE	Reject H_0 if $z \leqslant -1.645$; accept H_0 otherwise.

6.4 PERFORMING THE TEST

After the test procedure has been set up in accordance with Steps 1–3 above, it remains only to take the sample, compute the resulting value of the test statistic, and make a decision in accordance with the decision rule.

In the example used for the discussion above, an actual sample of 25 boxes yielded $\bar{y} = (298.71/25) = 11.948$. This value falls in the critical region ($\bar{y} \leqslant 11.97$) and so we must reject H_0. In terms of the critical z region ($z \leqslant -1.645$) we would state:

From the sample data, yielding $\bar{y} = 11.948$, we have

$$z = \frac{11.948 - 12}{\sqrt{.0004}} = \frac{-0.052}{0.02} = -2.6.$$

Reject H_0.

There is a difference of opinion among experimental scientists, and among statisticians, as to whether *accepting* H_0 is the logical alternative to *rejecting* H_0. The argument is that rejection is reasonable when an experimental result seems beyond the limits of chance, but acceptance should require much more than having the results of a single sample be better than improbable. If one is ruled by this argument, he replaces "accept H_0" by "nonreject H_0," "do not reject H_0," or "suspend judgment."

In a mathematical sense, this is only a matter of semantics. The decision rule is a two-decision rule: reject H_0 or take some other action. "Accept H_0" came into usage as the most obvious opposite of "reject H_0," but other terms will fit the test procedure satisfactorily. We shall consistently use *reject* H_0 and *accept* H_0 as the alternatives, but in both cases we shall extend the statement to include some assertion about "significance."

The test procedure that we have outlined is called both *test of hypothesis* and *test of significance*. It is the same test under either name. As a test of hypothesis, it tells us how to decide whether or not to reject a null hypothesis. As a test of significance, it tells us whether sample data are or are not significantly contrary to the null hypothesis. One's decision is most helpfully set forth if it is reported in both ways. Thus in our example above, we would report: Reject H_0; the observed sample mean box content is significantly less than 12 ounces, at the 5 percent level of significance.

Notice that the null hypothesis (H_0: $\mu = 12$) is either rejected or accepted (nonrejected); there can be no matter of "significance" involved in the statement of a hypothesized *population* value. On the other hand, a *sample* value (here $\bar{y} = 11.948$) is not a hypothetical one; it has actually been observed, and either is or is not *significantly* in line with H_A's alternative to H_0. Notice also that it is essential to state the level of significance since that is the criterion used to define significance.

6.5 THE DESCRIPTIVE LEVEL OF SIGNIFICANCE

There are many situations in which the experimental scientist insists that there is no reasonable way to choose a value for α, the level of significance. Without an α value, he of course cannot set up a decision rule in the manner of the foregoing test procedure, since he has no criterion for "significance." Such investigators prefer to calculate the value of the test statistic from the sample data and then ask for the probability that a value "as excessive" as that one would have occurred if H_0 were true.

In our example of box weight we would have the following:

$$\bar{y} = 11.948, \qquad z = -2.6;$$
$$P(\bar{Y} \leqslant 11.948 \mid H_0) = P(Z \leqslant -2.6) = 0.0047.$$

This tells us that the probability is about 0.5 percent that a sample mean as far below 12 as the one we got would occur if in fact the mean μ were 12. It thus argues strongly against H_0.

Such a probability is customarily labeled P and is reported so that the investigator or reader of the report can decide for himself whether the value is small enough to justify the judgment that the sample result was so unlikely, if H_0 were true, that H_0 must be rejected. P is thus the smallest value of α at which the given test statistic could be ruled significant. It is thus a kind of ex post facto level of significance, and has been given the name *descriptive level of significance:*

$$
\begin{aligned}
P &= \text{descriptive level of significance} \\
&= P(\text{test statistic value as excessive as the one} \\
&\qquad \text{observed} \mid H_0)
\end{aligned}
\tag{6.5.1}
$$

Even when a customary level of significance, like 5 percent or 1 percent, is used in a test of hypothesis, it is useful to report the P-value as additional information about the sample.

6.6 ONE-TAILED AND TWO-TAILED TESTS

In our example we had H_0: $\mu = 12$ versus H_A: $\mu < 12$. Here the alternative hypothesis is a one-sided alternative to H_0: just those values less than 12 are of interest to the test. As we saw, this led to a critical region of z values in the left-hand tail of the standard normal distribution. In this sense the test is a *one-tailed* test. Had our interest centered on values larger than 12, we would have had H_0: $\mu = 12$ versus H_A: $\mu > 12$, and again the test would have been one-tailed, this time having the critical region in the right-hand tail of the z distribution.

When H_A is a two-sided alternative to H_0, as in the case of H_0: $\mu = 12$ versus H_A: $\mu \neq 12$, there are two tails of interest in the distribution of the test statistic, and the critical region is composed of two parts, one in each tail. In the absence of any special considerations, the size α is divided evenly between the two tails, $\alpha/2$ for each of the two parts of the critical region.

Example 6.6.1

To make a convenient comparison of the one-tailed and two-tailed tests, let us go back to our example of weight of contents in boxes marked "12 ounces," and now set down the complete procedure of a *two-tailed* test, using the same (5 percent) level of significance.

H_0: $\mu = 12$ versus H_A: $\mu \neq 12$. Y is $N(\mu, 0.1)$; random sample, $n = 25$; $\alpha = .05$.

If H_0 is true, then

$$\frac{\bar{Y} - 12}{\sqrt{\dfrac{.01}{25}}} = Z.$$

Reject H_0 if $z \leq -1.960$ or if $z \geq +1.960$; accept H_0 otherwise.
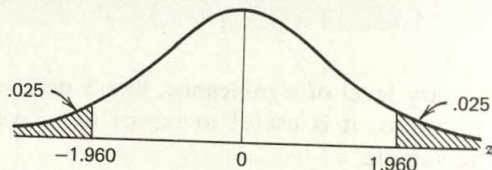
From the sample data (giving $\bar{y} = 11.948$) we have

$$z = \frac{11.948 - 12}{\sqrt{.01/25}} = \frac{-0.052}{\sqrt{.0004}} = \frac{-0.052}{0.02} = -2.60.$$

Reject H_0. The observed sample mean box content is significantly different from 12 ounces, at the 5 percent level of significance.

$$P = P(Z \leq -2.60 \quad \text{or} \quad Z \geq +2.60) = 2(.0047) = .0094.$$

Two things about the conclusion of the test should be given special notice:

1. The statement about significance is related to the alternative hypothesis H_A. Here H_A specifies $\mu \neq 12$. Hence the "significance" concerning the sample mean $\bar{y}$ has to do with $\bar{y}$ being *different from* 12, whether it is greater or less than 12 being beside the point. The critical region treats both alternatives equally so that the important characteristic at issue is *difference from* 12; we must state our conclusion consistently with this.

2. The *observed* statistic value will be in one tail or the other, but the P value must take into account two tails whenever the critical region involves two tails of a distribution. This is the reason why the P-value is defined as the probability of obtaining a statistic value "as excessive as" the one observed; in a two-tailed test this means "as far out, *in either direction*, as the one observed."

EXERCISES

6.6.1 In each of the following cases, consider that there is a random variable Y under study, that its standard deviation σ is taken as known, and that it is desired to test a hypothesis concerning its mean μ. Set up the test of the specified hypothesis, given the related facts as indicated.
a. $\mu = 50$ versus $\mu \neq 50$, given $\sigma = 10$, $n = 25$, $\alpha = .05$.
b. $\mu = 75$ versus $\mu > 75$, given $\sigma = 8$, $n = 50$, $\alpha = .01$.
c. $\mu = 300$ versus $\mu < 300$, given $\sigma = 20$, $n = 16$, $\alpha = .05$.

6.6.2 In a certain normal population σ is known to be 8.
a. Set up the test (that is, give the first three steps in the standard test procedure) for testing H_0: $\mu = 50$ versus H_A: $\mu > 50$, at the 5 percent level of significance, using sample size $n = 16$.
b. Notice that the critical region $z \geq z_c$ (where z_c is the proper critical value of z) can be expressed in the form $\bar{y} \geq k$ (by making use of the fact stated in Step 2 of the test). Find this form in the above case, and then express your decision rule in the form "Reject H_0 if $\bar{y} \geq$ ———; accept H_0 otherwise." Now what is the probability that you will *accept* H_0 if the population mean μ is actually 55?

6.6.3 Suppose that with a certain intelligence test the distribution of I.Q.s is known to be normal with standard deviation 16. Suppose also that there is a policy to conduct a special study of the school program in any group where the mean I.Q. is less than 90. A class of 36 students from a certain group is to be tested, and on the basis of the results a decision is to be made as to whether the special study should be instituted. Set up the appropriate test of hypothesis, taking significance level .01. What does this procedure assume about the relationship between the class and the group from which it came? What is the probability of deciding in favor of the special study if the group mean is really 80?

6.6.4   The life of a pair of army shoes is stated to be normally distributed with mean $\mu = 12$ months, and standard deviation $\sigma = 2$ months. The supply sergeant for company A in the 198th infantry regiment kept records on the life of army shoes in his company. He had to replace 100 pairs of shoes last year. When he calculated the average life he found it to be $\bar{y} = 11.7$ months. Using $\alpha = 5$ percent, is this sufficient evidence for the sergeant to complain to the supply officer about the quality of the shoe wear? Draw your own conclusions using your own choice of α and state justification for any actions you recommend.

6.6.5   The diameters of certain shafts must be less than 1.500 inches to be usable. The shafts are produced by a process that gives a normal distribution with a mean diameter $\mu = 1.490$ inches and a standard deviation $\sigma = .005$ inches. A metallurgical company using these shafts registered a complaint with the supplier. The company said that the average shaft they'd been receiving was $\bar{y} = 1.495$ inches based on a sample of $n = 49$ shafts and they had to reject too many shafts. Using $\alpha = 5$ percent, do you agree with the company's complaint? Why or why not?

6.7 TESTS CONCERNING μ WHEN σ IS UNKNOWN

The logic of every statistical test of hypothesis is precisely the same as that discussed above. Each test has the Steps 1–3 of Section 6.3, and the conclusion is carried out in the manner of Section 6.4. What changes from test to test is the specific detail of the steps. For any given set of conditions in Step 1, it is the role of the mathematical statistician to discover optimum procedures for Steps 2 and 3. The procedures which we present in this book have been thus worked out by statistical theorists, and confirmed in usefulness by applied practice in diverse fields of investigation.

As in the case of estimation, the most important test situation involving a population mean is the one in which the standard deviation σ is unknown also. Here we need only apply the relation (5.7.2) in Step 2 and then proceed with a *t-test* in the same way we used the *z*-test above.

Example 6.7.1

There is concern about strontium-90 radioactivity in milk. Measurement of such radioactivity is in terms of the number of picocuries of radiation per liter.

A certain standard sets 5 picocuries per liter as an acceptable limit for a milk-supply mean. A given milk supply is checked by use of a random sample of 10 units taken from the supply. The concern is whether the supply mean exceeds the 5 picocurie standard, and a 1 percent level of significance is used. In one test the sample gave the following results: 7.9, 9.1, 9.8, 8.4, 10.1, 7.6, 8.2, 9.9, 10.2, and 11.0. We assume the radioactivity per unit to be normally distributed.

H_0: $\mu = 5$ versus H_A: $\mu > 5$. Y is normal; random sample, $n = 10$; $\alpha = .01$.

If H_0 is true, then

$$\frac{\bar{Y} - 5}{\sqrt{s^2/10}} = t \text{ with 9 d.f.}$$

Reject H_0 if $t \geqslant 2.821$; accept H_0 otherwise.

Probability density function of t with 9 d.f.

.01

0   2.821
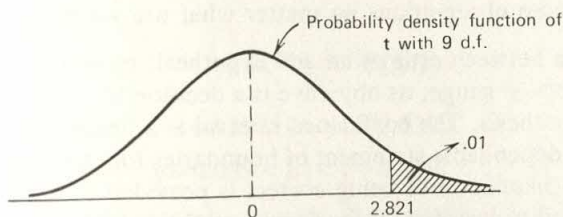
From the sample data we have the following.

Y	y²
7.9	62.41
9.1	82.81
9.8	96.04
8.4	70.56
10.1	102.01
7.6	57.76
8.2	67.24
9.9	98.01
10.2	104.04
11.0	121.00
92.2	861.88

$$\bar{y} = \frac{92.2}{10} = 9.22,$$

$$s^2 = \frac{861.88 - \dfrac{(92.2)^2}{10}}{9} = \frac{861.88 - \dfrac{8500.84}{10}}{9}$$

$$= \frac{861.88 - 850.08}{9} = \frac{11.80}{9} = 1.31.$$

$$t = \frac{9.22 - 5}{\sqrt{1.31/10}} = \frac{4.22}{\sqrt{0.131}} = \frac{4.22}{0.362} = 11.66.$$

Reject H_0. The observed sample mean radioactivity count is significantly greater than 5 picocuries per liter, at the 1 percent level of significance.

$$P < 0.0005.$$

The P value is exactly

$$P = P(t \geqslant 11.66 \mid 9 \text{ d.f.}).$$

In Table A-4 we see that 11.66 is beyond the largest entry for 9 d.f. (4.781) and that largest entry is the 99.95th percentile. Thus $P(t \geqslant 4.781 \mid 9 \text{ d.f.}) = .0005$, and so, since $P(t \geqslant 11.66 \mid 9 \text{ d.f.}) < P(t \geqslant 4.781 \mid 9 \text{ d.f.})$, we must have $P < .0005$.

6.8 RELATION BETWEEN TESTING AND ESTIMATING

Each of the above examples of test of hypothesis concerning a population mean μ shows that we use the same statistic and the same probability distribution theory as in making a confidence interval estimate of μ. This relationship is true in general for estimating a parameter on the one hand and testing a hypothesis concerning it on the other. The circumstance is of course not surprising, since a given chance situation involves a single probability structure and yields the same observations no matter what use we intend for the data.

The distinction between estimation and hypothesis testing is one of purpose. The test is a go-no-go gauge; its objective is a decision to reject or not to reject a stated null hypothesis. The confidence interval is a bracket for a parameter; its objective is a dependable statement of boundaries for the actual value of the parameter. High likelihood of being correct is provided in each procedure by use of the probability structure of the sample statistic, controlling at a low level (the level of significance) the probability of falsely rejecting a null hypothesis, controlling at a high level (the confidence coefficient) the probability of constructing a correct interval for the parameter value.

Consider the situation in Example 6.6.1. There we had a test of H_0: $\mu = 12$ against the alternative H_A: $\mu \neq 12$, and rejected H_0 at the 5 percent level of significance, on the basis of a sample mean $\bar{y} = 11.948$. An obvious question is: If you believe μ is not 12, what value do you believe μ does have? The answer can be given by a confidence interval. The limits of a 95 percent confidence interval are

$$11.948 \pm 1.960 \sqrt{\frac{.01}{.25}}$$

$$= 11.948 \pm 1.960\sqrt{.0004}$$
$$= 11.948 \pm 1.960(.02)$$
$$= 11.948 \pm 0.039,$$

so that a 95 percent confidence interval for μ is

$$11.909 < \mu < 11.987. \tag{6.8.1}$$

We can notice that this interval does not include the value 12. Hence $\mu = 12$ is outside the boundaries of our 95 percent confidence, and this is consistent with our having rejected H_0: $\mu = 12$ at the 5 percent level of significance.

Such a relationship is no mere coincidence. The confidence-limit value of z for a 95 percent confidence interval is precisely the same (1.96) as the critical value of z for a two-tailed test at the 5 percent level of significance. The decision rule in the test was: reject H_0 if $z \leqslant -1.960$ or $z \geqslant +1.960$; accept H_0 otherwise. This makes the *acceptance* region

$$-1.960 < z < +1.960,$$

and, in terms of $\bar{y}$ and a hypothetical value μ_0 for the mean μ, that acceptance region is

$$-1.960 < \frac{\bar{y} - \mu_0}{.02} < 1.960,$$

$$-0.039 < \bar{y} - \mu_0 < 0.039,$$

$$\bar{y} - 0.039 < \mu_0 < \bar{y} + 0.039. \tag{6.8.2}$$

Thus any hypothetical value of μ *outside* the limits in (6.8.2) would be *rejected* in a two-tailed test at the 5 percent level of significance. But the limits in (6.8.2) are precisely the limits of a 95 percent confidence interval for μ, as we saw in obtaining (6.8.1).

In general, a c percent confidence interval for a population parameter contains all of the values that could be accepted in a two-tailed test of hypothesis about that parameter at the $(100 - c)$ percent level of significance. There is a similar relationship between one-tailed tests and *one-sided* confidence intervals but these are not considered in this book.

If the investigator wants to pursue analysis beyond just a decision to reject or not reject a null hypothesis, he customarily starts with a confidence interval. This is especially pertinent when a null hypothesis is rejected. Many times the test of hypothesis is a screening process to uncover "significant" cases for more detailed study.

A confidence interval following a test of hypothesis is usually the two-boundary interval that we have studied, even when the test is one-tailed. In such cases one must bear in mind that the confidence-limit value of z or t is *not* the same as the critical value in the test. The following cases from our earlier examples illustrate the point.

Example 6.8.1

Our first example, introduced in Section 6.2 and carried along through the various test steps, involved H_0: $\mu = 12$ versus H_A: <12. The test was conducted at the 5 percent level of significance, with the critical region being $z \leqslant -1.645$. We observed $\bar{y} = 11.948$ and rejected H_0. A 95 percent confidence interval to follow this test would be exactly the same one (6.8.1) that we constructed after the two-tailed test using the same level of significance and the same value of $\bar{y}$.

Example 6.8.2

In Example 6.7.1 we had a t test of H_0: $\mu = 5$ versus H_A: $\mu > 5$. At the 1 percent level of significance the critical region was $t \geqslant 2.821$, the critical value being the 99th percentile of t with 9 d.f. For a 99 percent confidence interval for μ we must proceed as specified by (5.7.3), using data in Example 6.7.1 to obtain the limits

$$9.22 \pm (3.250)(0.362)$$

$$= 9.22 \pm 1.18,$$

thus obtaining the 99 percent confidence interval

$$8.04 < \mu < 10.40.$$

EXERCISES

[Consider it satisfactory to assume that Y is essentially normally distributed in all the following exercises.]

6.8.1 In testing H_0: $\mu = 3.5$ against H_A: $\mu > 3.5$, in a normal population, a sample of size 20 yields $\bar{y} = 3.591$ and $s = 0.1527$. What decision should be made at the 1 percent level of significance?

6.8.2 For what values of $\bar{y}$ would you reject the null hypothesis in a two-sided test for each of the following situations? (a) H_0: $\mu = 0.8$, given $\sigma = 0.2$, $n = 16$, $\alpha = .05$, (b) H_0: $\mu = 0.8$, $s = 0.2$, $n = 16$, $\alpha = .05$, (c) H_0: $\mu = 10$, $\sigma = 2$, $n = 100$, $\alpha = .10$, (d) H_0: $\mu = 10$, $s = 2$, $n = 100$, $\alpha = .10$, (e) H_0: $\mu = 40$, $\sigma = 5$, $n = 4$, $\alpha = .01$, and (f) H_0: $\mu = 40$, $s = 5$, $n = 4$, $\alpha = .01$.

6.8.3 Let us assume that regular accounting procedures turn up numerical errors averaging about $21.40 per month. A new method of accounting involving some automation is under investigation. We would like to determine whether the cost of numerical errors will change. We desire to run a risk of 5 percent of being wrong when we say that there has been a change in the average dollar error per month. A sample of $n = 20$ observations using the automated system was taken and the following calculations were made:

$$\bar{y} = \$21.50; \qquad s = \$0.40$$

a. Is there sufficient evidence to warrant taking action on a change in dollar errors per month? State the hypothesis to be tested. Calculate the statistic needed to test the hypothesis.

b. Suppose we are interested in making the change to the automated system only if there is a decrease in the average dollar error per month. Write down the hypothesis for this case. What is your decision for this case?

6.8.4 Each department in a metals-processing company inspects its finished product with its own micrometer. These micrometers are calibrated using "¼" standard gauge blocks, for which the true measurement should be .025. It has been suggested that these micrometers are all different. To test this hypothesis, eight identical standard gauge blocks were submitted to each department in a random order and were measured. Micrometer data from the first department were:

.0251
.0252
.0248
.0252
.0247
.0254
.0245
.0246

a. If you use an α-risk of 5 percent, do you agree or disagree with the suggestion that the department's micrometer needs calibrating?
b. The second department's micrometer readings were as follows:

.0251
.0249
.0254
.0253
.0249
.0248
.0248
.0254

Again using $\alpha = 5$ percent, do you agree or disagree with the suggestion that this department's micrometer needs calibration?
c. Using the confidence interval approach from Chapter 5, calculate 95 percent confidence limits on the true mean reading of block size for each micrometer. Can you say that the two true means are different?
d. Can you think of a better way to determine whether the two micrometers are different from each other?

6.8.5 A good T.V. commercial has a percent recall score of 40. A new T.V. commercial was given a trial in a southwestern U.S. city, and the following percent recall scores were obtained: 42, 18, 46, 35, 51, 29, 45, 45, 36, and 35.
a. With $\alpha = 10$ percent, use the t test to determine whether this T.V. commercial's percent recall is significantly less than the standard of 40.
b. What conclusions can you draw about this result?
c. How would you conduct an experiment like this if you were trying to learn how good the new T.V. commercial really is?

6.9 TESTS CONCERNING THE DIFFERENCE BETWEEN TWO POPULATION MEANS

In Sections 5.8 and 5.9 we considered the confidence-interval estimation of the difference $\mu_1 - \mu_2$ between two population means. In Section 5.8 we had the situation in which the population standard deviations σ_1 and σ_2 are known, and in Section 5.9 we treated the more practical situation in which σ_1 and σ_2 are unknown. In this latter situation we restricted ourselves to the case wherein $\sigma_1 = \sigma_2$. In the matter of hypothesis testing we shall limit ourselves to this restricted practical case.

Example 6.9.1

In Examples 6.7.1 and 6.8.2 we considered data on strontium-90 radioactivity in the milk of a certain given supply. A milk supply in a different region was checked by means of a random sample of 12 units measured for strontium-90 activity. The resulting measurements (in picocuries per liter) were 4.6, 5.1, 8.7, 6.9, 6.1, 5.0, 5.2, 5.0, 5.6, 5.2, 3.9, and 4.0. Are the observed means in the two regions significantly different, at the 1 percent level of significance?

Assuming the radioactivity per unit to be normally distributed in each region, and its standard deviation to be the same in both regions, we can apply the probability information of (5.9.2) and conduct the test of significance as follows:

H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 \neq \mu_2$. Y_1 and Y_2 normal, $\sigma_1 = \sigma_2$; independent random samples, $n_1 = 10$, $n_2 = 12$, $\alpha = .01$.

If H_0 is true, then

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{\sqrt{\dfrac{s_p^2}{10} + \dfrac{s_p^2}{12}}} = t \text{ with 20 d.f.}$$

Reject H_0 if $t \leqslant -2.845$ or if $t \geqslant +2.845$; accept H_0 otherwise.

Probability density function of t with 20 d.f.

.005   .005
−2.845   0   2.845
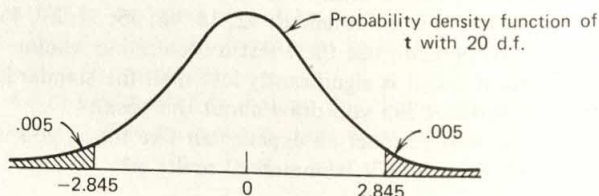
From the sample data we have the following:

From Example 6.7.1,

$n_1 = 10,$

$\bar{y}_1 = 9.22,$

$s_1^2 = \dfrac{11.80}{9} = 1.31$

y_2	y_2^2
4.6	21.16
5.1	26.01
8.7	75.69
6.9	47.61
6.1	37.21
5.0	25.00
5.2	27.04
5.0	25.00
5.6	31.36
5.2	27.04
3.9	15.21
4.0	16.00
65.3	374.33

$\bar{y}_2 = \dfrac{65.3}{12} = 5.44,$

$s_2^2 = \dfrac{374.33 - \dfrac{(65.3)^2}{12}}{11}$

$= \dfrac{374.33 - \dfrac{4264.09}{12}}{11}$

$= \dfrac{374.33 - 355.34}{11}$

$= \dfrac{18.99}{11} = 1.73.$

$$s_p^2 = \frac{11.80 + 18.99}{9 + 11} = \frac{30.79}{20} = 1.54,$$

$$t = \frac{(9.22 - 5.44) - 0}{\sqrt{\dfrac{1.54}{10} + \dfrac{1.54}{12}}} = \frac{3.78}{\sqrt{.154 + .128}} = \frac{3.78}{\sqrt{.282}} = \frac{3.78}{0.531} = 7.12.$$

Reject H_0. The observed mean concentrations in the two regions differ significantly, at the 1 percent level of significance.

Because the test is two-tailed, the descriptive level of significance P is

$$P = 2 \cdot P(t \geqslant 7.12 \mid 20 \text{ d.f.})$$

In Table A-4 we see that 7.12 is beyond the largest tabulated entry for 20 d.f. (3.850). Since that entry is the 99.95th percentile, the probability beyond it is 0.0005, and so $P(t \geqslant 7.12 \mid 20 \text{ d.f.})$ is less than 0.0005. Twice the probability is thus less than double the value 0.0005, and therefore we report:

$$P < 0.001.$$

The limits of a 99 percent confidence interval for the difference in population means $\mu_1 - \mu_2$ can be quickly set down here, since the numerical values needed in (5.9.3) have already been found and used in the test of significance:

$$3.78 \pm 2.845(0.531) = 3.78 \pm 1.51.$$

Thus a 99 percent confidence interval for the difference in population mean strontium-90 activity in milk (picocuries per liter) in the two regions is

$$2.27 < \mu_1 - \mu_2 < 5.29.$$

EXERCISES

[Consider it satisfactory to assume that random variables are normally distributed and $\sigma_1 = \sigma_2$ in each of the following exercises.]

6.9.1 Company A produces a special pharmaceutical tablet on a press. The tablet press has two outlets. From each outlet a total of 10 tablets is randomly selected, weighed, and recorded. Based on the results shown below, is there a statistically significant difference between the two outlets? (Use $\alpha = 5$ percent.)

Left Side		Right Side	
10.90	10.94	10.95	10.92
10.94	10.94	10.92	10.93
10.97	10.91	10.92	10.93
10.95	10.95	10.90	10.94
10.90	10.96	10.91	10.92

6.9.2 A new brand promotion method is to be tested in one sales area. The effectiveness of this method will be judged by the weekly shipments of the brand to this area, measured as a percentage of the 1959 base period. A second sales area, in which the normal promotion method is continued, will be used as a control. Because the different promotion methods probably involve different sales cycles, the weekly data from the two districts cannot be paired. The following data are collected over a 13-week period:

Week	Test District	Control District
1	109	135
2	108	101
3	127	111
4	136	119
5	101	106
6	111	117
7	136	106
8	109	91
9	119	88
10	126	92
11	85	105
12	145	98
13	147	96

a. On the basis of the above data, we wish to determine whether the new method is more effective than the old method. With an α risk of 0.05, could you say that the new method is more effective? [Note that H_A is $\mu_1 > \mu_2$, that is, $\mu_1 - \mu_2 > 0$.]

b. What is the best estimate of the change in shipments brought about by the new method? Give a 90 percent confidence interval for the true change.

6.9.3 All accounting information is available from two plants. From plant A, 10 random accounts were selected and checked for percent errors. From plant B, eight accounts were chosen at random. The results are as follows:

Plant A		Plant B	
12.1	12.4	13.6	12.5
12.0	10.6	12.4	13.6
12.1	11.8	11.8	12.8
11.8	12.4	11.9	13.2
11.5	11.9		

Would you consider these plants different? Why? (Choose your own α.)

6.9.4 Twenty new accounting men are to be trained in our accounting system. Two methods of training have been suggested. In order to test these methods, the trainees are divided into two groups; group 1 is trained by method A and group 2, by method B. After three months of training, a test set of material is given to all trainees. The scores are shown below.

Training Method A		Training Method B	
96	92	94	96
90	94	93	84
93	83	86	98
88	80	89	95
86	98	94	91

a. Write out the hypothesis to be tested.
b. Using $\alpha = 10$ percent, would you decide that one training method is better than the other?

6.9.5 A company is interested in the ability of denture cleansers to prevent the accumulation of stain on dentures. Two brands were compared. Thirty subjects with dentures were selected at random and fifteen were assigned at random to each brand. All dentures were thoroughly cleaned by the dental technician at the beginning of the test period. The quantity of stain was then measured at the end of a test period of 3 months. The following results were obtained.

Stain Scores

Brand A		Brand B	
8.0	7.5	5.0	6.5
10.0	10.0	6.0	8.0
7.5	17.0	11.0	8.5
5.5	15.5	9.0	6.5
8.0	9.5	6.0	4.0
3.0	10.0	5.0	8.0
11.0	11.5	8.0	7.5
8.5		6.0	

Using an α-risk of 5 percent, what is your conclusion about the relative merits of the two brands?

6.9.6 Two vendors have been asked to furnish random samples of "Supercleanser" cans made with a board that will minimize the moisture–vapor transmission rate. The following figures are in grams per 100 inches squared per 24 hours.

Supplier A	Supplier B
.047	.054
.047	.052
.055	.052
.053	.050
.049	.051
.047	.054
.051	
.046	

Code the data by multiplying by 1000 and subtracting 40.
a. Is there any difference in the means of cans supplied by the two suppliers? [Use $\alpha = .05$.]
b. Calculate a 95 percent confidence interval for the true difference between the means.
c. What is the relationship between the confidence interval and test of significance in this example?

6.9.7 Industrial engineers working with chemical engineers are responsible for making changes in process methods and in process equipment in order to improve performance. In a typical sulfonation process, an industrial engineer suggested making the nozzle orifices smaller in order to increase the percent completeness of the finished product. The following data were obtained before and after the change:

Before Change		After Change	
95.5	96.1	96.4	95.8
95.2	95.8	95.6	96.6
95.0	95.8	95.6	95.7
94.9	95.7	96.5	95.6
95.4	95.8	95.7	95.4
95.3	95.8	95.7	95.3
94.6	95.4	96.2	96.5
95.8	95.1	95.8	96.2
95.5	95.6	96.7	96.4
95.8		96.6	

Using a t test with $\alpha = 1$ percent, do you agree that a significant increase in percentage of completeness has occurred?

6.9.8 a. Using the following sample data, find a 95 percent confidence interval for the population mean:

$$y: 7, 5, 8, 6, 5, 9$$

b. Consider the following sample data taken from a second population:

$$y: 5, 6, 3, 2, 3, 4, 4, 5$$

Test the hypothesis that the population mean in this case and that in the case considered in (a) are equal. (Use 1 percent level of significance.)

c. Find a 99 percent confidence interval for the difference between the means of the two populations.

6.9.9 We wish to determine whether women will use more of a certain dish-washing liquid if the size of the cap on the container is increased. A random sample of 16 women was given a container with one cap size to use for a specified number of dish washings. Another random sample of 16 women was given a container with the other cap size to use for an equal number of dish washings. The containers with the larger caps are marked "K" and the containers with small caps, "J." The following total usage data were obtained:

Usage with J (cubic centimeters)	Usage with K (cubic centimeters)
55	58
75	74
45	47
96	104
38	38
62	66
55	57
82	80
75	79
43	49
54	57
31	39
44	47
55	58
53	51
69	71

a. Does the larger cap result in an increased usage? (Use an α risk of 0.05.)

b. How would you improve this testing procedure?

6.10 TESTS CONCERNING THE BINOMIAL PROPORTION p

We can carry forward the estimation discussion of Section 5.10 and apply our standard hypothesis-testing procedure to obtain tests about p. The essential probability theory is stated by the relation (5.10.3), so that, if H_0 hypothesizes that p has a certain specific value, say p_0, then *under H_0* we have

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} \approx Z \quad \text{if } np_0 > 5 \text{ and } nq_0 > 5, \quad \text{where} \quad q_0 = 1 - p_0.$$

From this we proceed as in our z tests concerning a mean μ. (Note that the difficulty we had in Section 5.10, where we had to use $\hat{p}$ in the standard error, does not arise here, since p_0 is hypothesized and used wherever needed.)

Example 6.10.1

A random sample of 150 voters drawn from a certain town is questioned for opinion about a town ordinance under consideration. Favorable opinion is given by 81. Is this proportion significantly less than 60 percent, at the 1 percent level of significance?

$$H_0: p = .60 \text{ versus } H_A: p < .60. \quad \begin{array}{l} p = \text{binomial proportion;} \\ \text{random sample, } n = 150; \ \alpha = .01. \end{array}$$

If H_0 is true, then

$$\frac{\hat{p} - .60}{\sqrt{\frac{(.60)(.40)}{150}}} \approx Z.$$

Reject H_0 if $z \le -2.326$; accept H_0 otherwise.

$$\hat{p} = \frac{81}{150} = 0.54,$$

$$z = \frac{0.54 - 0.60}{\sqrt{\frac{(.60)(.40)}{150}}} = \frac{-0.06}{\sqrt{\frac{.2400}{150}}} = \frac{-0.06}{\sqrt{.0016}} = \frac{-0.06}{0.04} = -1.50.$$

Accept H_0. The observed sample proportion of voters in favor of the proposed ordinance is not significantly less than 60 percent, at the 1 percent level of significance.

$$P = P(Z \le -1.50) = 0.0668.$$

If in addition a confidence interval for p is desired, using Section 5.10, a 99 percent confidence interval would have limits:

$$.54 \pm 2.58 \sqrt{\frac{(.54)(.46)}{150}} = .54 \pm 2.58 \sqrt{\frac{.2484}{150}}$$

$$= .54 \pm 2.58 \sqrt{.00166} = .54 \pm (2.58)(.0407)$$

$$= .54 \pm .105,$$

giving the 99 percent confidence interval: 43.5 percent $< p <$ 64.5 percent

Example 6.10.2

Refer to the student data given in Chapter 1, in particular the opinion on whether marijuana should be legalized. Responses 1 ("strongly disagree") and 2 ("mildly disagree") indicate opinion against legalization. At the 5 percent level of significance, test the hypothesis that the population from which the sample was drawn has 25 percent of its members unfavorable to legalizing marijuana, as against some other percentage.

H_0: $p = .25$ versus H_A: $p \neq .25$. $p =$ binomial proportion;
random sample, $n = 180$; $\alpha = .05$.

If H_0 is true, then

$$\frac{\hat{p} - .25}{\sqrt{\frac{(.25)(.75)}{180}}} \approx Z.$$

Reject H_0 if $z \leq -1.960$ or if $z \geq +1.960$; accept H_0 otherwise. From the sample data, giving 63 replies in categories 1 or 2, we have

$$\hat{p} = \frac{63}{180} = 0.35,$$

$$z = \frac{0.35 - 0.25}{\sqrt{\frac{(.25)(.75)}{180}}} = \frac{0.10}{\sqrt{\frac{.1875}{180}}} = \frac{0.10}{\sqrt{.00104}} = \frac{0.10}{0.0322} = 3.11.$$

Reject H_0. The observed sample proportion of students unfavorable to legalizing marijuana is significantly different from 25 percent, at the 5 percent level of significance.

Using the closest z value (3.090) in Table A-3, we have

$$P = P(Z \geq 3.11 \text{ or } Z \leq -3.11) = 2(.0010) = .0020.$$

6.11 TESTS CONCERNING THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS

In Section 5.12 we worked out a procedure for constructing an approximate confidence interval for the difference $p_1 - p_2$ between two population proportions. For this we used the relation (5.12.3) as a starting point and then introduced further approximation by using estimates for p_1 and p_2 in the standard error, thus taking the probability structure as

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1 \hat{q}_1}{n_1} + \dfrac{\hat{p}_2 \hat{q}_2}{n_2}}} \approx Z \qquad \begin{array}{l} \text{if } n_1 p_1, \ n_1 q_1, \ n_2 p_2, \ n_2 q_2 \\ \text{are all greater than 5.} \end{array} \qquad (6.11.1)$$

In the customary test of "the significance of the difference between (sample) proportions," the null hypothesis H_0 specifies *no difference*; H_0: $p_1 = p_2$ or $p_1 - p_2 = 0$. Under this hypothesis the estimates $\hat{p}_1$ and $\hat{p}_2$ are both estimates of the same parameter value, say p, the common value of p_1 and p_2 under the hypothesis that $p_1 = p_2$. In this circumstance $\hat{p}_1$ and $\hat{p}_2$ should be pooled to give a single estimate of the single value p, just as we pooled s_1^2 and s_2^2 when we assumed $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Again we use a weighted average, this time using sample size as weight. This leads to a particularly simple and sensible pooled estimate:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{\text{Total number of "heads" in both samples}}{\text{Total number of observations in both samples}} \qquad (6.11.2)$$

This estimate replaces $\hat{p}_1$ and $\hat{p}_2$ in (6.11.1) when we are operating under the hypothesis that $p_1 = p_2$. That hypothesis also causes $p_1 - p_2$ to be zero, and so the modification of (6.11.1) gives us the following probability relation to use in our test of hypothesis.

Under H_0: $p_1 = p_2$,

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}\hat{q}}{n_1} + \dfrac{\hat{p}\hat{q}}{n_2}}} \approx Z, \qquad \hat{q} = 1 - \hat{p}. \qquad (6.11.3)$$

The approximation is considered satisfactory if $n_1\hat{p}$, $n_1\hat{q}$, $n_2\hat{p}$, $n_2\hat{q}$ are all greater than 5.

Example 6.11.1

Test the significance of the difference between males and females as to unfavorable opinion on legalizing marijuana, according to the sample data in Chapter 1. Use the 5 percent level of significance.

H_0: $p_F = p_M$ versus H_A: $p_F \neq p_M$. p_F = binomial proportion, females;
p_M = binomial proportion, males;
independent random samples,
$n_F = 76$, $n_M = 104$; $\alpha = .05$.

If H_0 is true, then

$$\frac{\hat{p}_F - \hat{p}_M}{\sqrt{\dfrac{\hat{p}\hat{q}}{76} + \dfrac{\hat{p}\hat{q}}{104}}} \approx Z.$$

Reject H_0 if $z \leqslant -1.960$ or if $z \geqslant +1.960$; accept H_0 otherwise. From the sample data we have

$$\hat{p}_F = \frac{32}{76} = 0.421, \qquad \hat{p}_M = \frac{31}{104} = 0.298, \qquad \hat{p} = \frac{32+31}{76+104} = \frac{63}{180} = 0.350$$

$$z = \frac{0.421 - 0.298}{\sqrt{\dfrac{(.350)(.650)}{76} + \dfrac{(.350)(.650)}{104}}} = \frac{0.123}{\sqrt{\dfrac{.2275}{76} + \dfrac{.2275}{104}}}$$

$$= \frac{0.123}{\sqrt{.00299 + .00219}}$$

$$= \frac{0.123}{\sqrt{.00518}} = \frac{0.123}{0.0720} = 1.71.$$

Accept H_0. The observed difference between male and female students as to sample proportion unfavorable to legalizing marijuana is not significant at the 5 percent level.

$$P \approx P(Z \geqslant 1.7 \text{ or } Z \leqslant -1.7) = 2(.0446) = .0892.$$

If we want a confidence interval for the difference $p_F - p_M$ we must use (5.12.4), keeping $\hat{p}_1$ and $\hat{p}_2$ distinct in the standard error rather than using pooled $\hat{p}$ as in the above test. The test proceeds under a null hypothesis that $p_F = p_M$; such a hypothesis is clearly not applicable when we seek bounds on a nonzero difference. The limits of an approximate 95 percent confidence interval for $p_F - p_M$ are

$$(0.421 - 0.298) \pm 1.96 \sqrt{\frac{(.421)(.579)}{76} + \frac{(.298)(.702)}{104}}$$

$$= 0.123 \pm 1.96 \sqrt{\frac{.2438}{76} + \frac{.2092}{104}}$$

$$= 0.123 \pm 1.96 \sqrt{.00321 + .00201}$$

$$= 0.123 \pm 1.96 \sqrt{.00522}$$

$$= 0.123 \pm (1.96)(.0722)$$

$$= 0.123 \pm 0.142,$$

so that an approximate 95 percent confidence interval is

$$-0.019 < p_F - p_M < +0.265$$

or

$$-1.9 \text{ percent} < p_F - p_M < +26.5 \text{ percent}.$$

Notice that the confidence bounds extend from where the female proportion is 1.9 percentage points *lower than* the male to where it is 26.5 percentage points *higher*. This is of course consistent with being *unable* to reject a null hypothesis of no difference, since the value $p_F - p_M = 0$ lies within the confidence interval.

EXERCISES

6.11.1 Suppose that for a certain disease the mean mortality is 36 out of 100 attacks. If under a new treatment there are 120 deaths out of 400 attacks, what would you say, at the 1 percent level, about the new treatment?

6.11.2 The probability of winning a game of "craps" is 0.495. Suppose a player wins 60 games out of 100. At the 1 percent level, would you consider this a significant deviation from the expected? What would be your decision at the 5 percent level of significance?

6.11.3 In a test of effectiveness, 200 insects were sprayed with insecticide A and 300 were sprayed with insecticide B. The numbers of insect deaths were 150 and 210, respectively. At the 5 percent level, is there a significant difference in effectiveness between the two insecticides?

6.11.4 With the standard process of manufacturing a certain article, 5 percent of the produced units are defective. A new time- and money-saving process will be installed if it does not significantly increase this proportion of defectives. A test run of 900 units produced by the new process shows 55 defective. At the 5 percent level of significance, what conclusion would you draw?

6.11.5 It is believed that 20 percent of the voters in a certain community are Independent voters. A poll is taken of 196 voters constituting a representative sample. Of these, 27 state that they are Independent. Does this result support the belief at the 5 percent level of significance?

6.11.6 Referring to the mean proportion of deaths given in Exercise 4.9.2, consider a group of 1000 persons of age 60 about whom there is a strong presumption of unusually good health. If 200 of this group die before age 70, what conclusion (at the 1 percent level of significance) could you draw?

6.11.7 A random sample of $n = 300$ housewives was given two products, A and B, to use. One week later each housewife was asked which product she preferred. Fifty-five percent preferred A. Is this sufficient evidence with $\alpha = 5$ percent to conclude that A is truly preferred to B?

6.11.8 In a poll of 100 Independent voters in a local area, 40 percent stated that they would vote for candidate A, 48 percent said that they would vote for candidate B, and 12 percent said that they were undecided. Assuming the 12 percent will vote 6 percent for A and 6 percent for B, is this enough evidence for you to say B will win the election? Why or why not?

6.11.9 A test to determine whether a new deodorant (D) is preferred to a standard well-known brand (S) was conducted in two different cities. In city A, 52 percent of the $n = 200$ people tested said that they preferred brand D, the new brand. In city B, 55 percent of the $n = 200$ people tested said that they preferred brand D.

    a. Using $\alpha = 10$ percent and these data, is there sufficient evidence to state that the preference in city B for product D is significantly higher than in city A?

    b. Overall, how many people preferred product D? If we pool all of the data, what proportion of people prefer product D?

    c. Using the pooled data and $\alpha = 5$ percent, test the hypothesis that the products S and D are equally preferred. What conclusion do you make?

    d. Comment on the assumptions one makes by pooling these results from the two different cities.

7.1 INTRODUCTION

There are many situations when qualitative or "nominal" data are collected and form the major evidence for or against certain conjectures or hypotheses. One of the most useful techniques employed for analyzing these data is the chi-square test of hypothesis. In this chapter, we will cover only one of the more elementary topics that utilize the chi-square (χ^2) distribution; namely, the $r \times c$ contingency table. For those interested in other uses of chi-square tests, a brief outline of other topics has been included.

7.2 A BINOMIAL PROBLEM

Let us consider another simple application of the binomial distribution. A coin is tossed 100 times and the following results (Table 7.2.1) are obtained:

TABLE 7.2.1

Heads	Tails	Total
60	40	100

As a result of this experiment, is there sufficient evidence to state that the coin is biased, that is, more apt to give heads than tails, or vice-versa? This question can be precisely stated in the form of a test of the hypothesis:

$$H_0: \; p_H = p_T = .50$$

7

Sorting Out
The
Categories

7

TELL IT LIKE IT IS*

"I'M 56% IN AGREEMENT, 31% IN
DISAGREEMENT, AND 13% UNDECIDED."

*By Ralph Dunagin. Courtesy of Field Newspaper Syndicate.

versus the two-tailed alternative:

$$H_A: \ p_H \neq .50$$

Let us prespecify the probability of rejecting the hypothesis of a fair coin when we should not do so at $\alpha = 5$ percent. The binomial-distribution approach to this problem is to calculate the observed proportion of heads from the sample ($\hat{p}_H$) and use the normal approximation to the binomial (since $np_0 > 5$) to test the hypothesis H_0. We calculate:

$$z = \frac{\hat{p}_H - .50}{\sqrt{\dfrac{(.50)(.50)}{100}}}$$

and compare this calculated z-value with the value of Z for the normal distribution that cuts off 2.5 percent of the area in each tail of the distribution. Using the table of the normal distribution (Table A-3), we find this to be $z_* = 1.96$.

In this example, the observed proportion and the calculated z are shown as follows:

$$z = \frac{.6 - .5}{\sqrt{\dfrac{(.50)(.50)}{100}}} = \frac{.1}{\sqrt{.0025}} = \frac{.1}{.05} = 2.00$$

Since the calculated value of $z = 2.00$ is greater than the critical value of $z_* = 1.96$, the experimenter concludes that the coin is biased, rejects H_0, and accepts H_A.

7.3 1×2 TABLES

Another approach to the solution of this type of problem is through the use of the more generally applicable χ^2 distribution. The χ^2 statistic is defined as follows:

$$\chi_l^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},\qquad (7.3.1)$$

where o_i = observed frequency in the ith category
  e_i = expected (theoretical) frequency in the ith category
  k = number of categories into which the data are divided
  χ_l^2 = a random variable that has the χ^2
    distribution with l degrees of freedom.

The probability pattern of the χ^2 statistic was worked out by Karl Pearson and R. A. Fisher in the early 1900s. As in the case of Student's t, there is a different distribution for every different number of degrees of freedom. Tables of these distributions are readily available (e.g., Table A-5). Figure 7.3.1 graphs a few illustrative cases.

Let us consider our binomial problem in terms of χ^2. Under the hypothesis that the coin is unbiased, the expected number of heads in 100 tosses of the coin is $\mathscr{E}(Y)$, where Y has the binomial distribution with $n = 100$ and $p = p_0 = .50$. Thus the expected frequency in the first category (heads) is $100(.50) = 50$, and then the following table (Table 7.3.1) can be constructed:
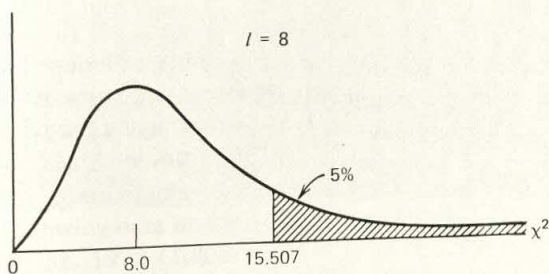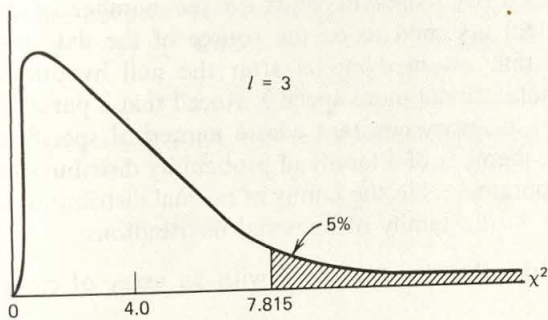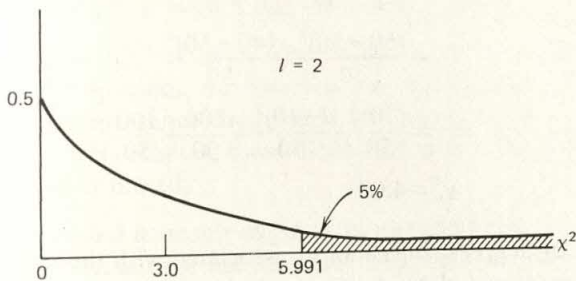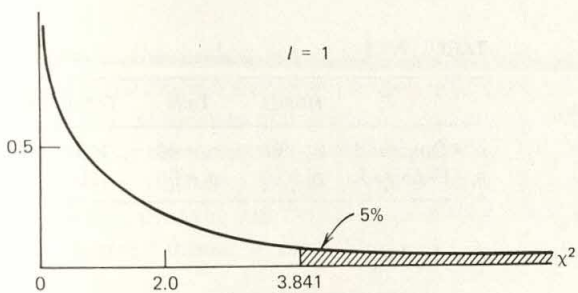
FIGURE 7.3.1 χ^2 distribution with *l* degrees of freedom.

TABLE 7.3.1

	Heads	Tails	Total
o_i = Observed	$o_1 = 60$	$o_2 = 40$	100
e_i = Expected	$e_1 = 50$	$e_2 = 50$	100

Using these values, we calculate

$$\chi_i^2 = \sum_{i=1}^{k=2} \frac{(o_i - e_i)^2}{e_i}$$

$$= \frac{(60 - 50)^2}{50} + \frac{(40 - 50)^2}{50}$$

$$= \frac{(10)^2}{50} + \frac{(-10)^2}{50} = \frac{100}{50} + \frac{100}{50}$$

$$\chi_i^2 = 4.00$$

The number of degrees of freedom l associated with the random variable χ_i^2 in tests on categorized data depends on the number of parameters in the probability model assumed to be the source of the data and the number of parameters in that assumed model after the null hypothesis H_0 makes the underlying general model more specific. Recall that a parameter in a probability model is an arbitrary constant whose numerical specification helps to pin down a specific member of a family of probability distributions. For example, μ and σ are the parameters in the family of normal distributions, and n and p are the parameters in the family of binomial distributions.

When χ_i^2 is the statistic associated with an array of categorized data, l is defined as

$$l = \begin{pmatrix} \text{Number of independent} \\ \text{parameters in the assumed} \\ \text{underlying model} \end{pmatrix} - \begin{pmatrix} \text{number of independent} \\ \text{parameters in the assumed} \\ \text{underlying model as} \\ \text{modified by } H_0 \end{pmatrix} \quad (7.3.2)$$

In our coin example, the underlying model (for the first part of Formula 7.3.2) is the binomial distribution with $n = 100$ and general p. In that model the parameters are p and q, subject to the restriction $p + q = 1$. Thus there is one *independent* parameter in the assumed underlying model. For the second part of Formula (7.3.2), the null hypothesis specifies that the distribution has $n = 100$ and also $p = 0.5$; thus the number of parameters under H_0 that are left unspecified in the assumed model is zero. Thus

$$l = 1 - 0 = 1,$$

and the χ^2 statistic is $\chi_1^2 : \chi^2$ with 1 d.f.

Referring to Figure 7.3.1 of the χ^2 distribution with $l = 1$ degree of freedom, using Table A-5 for specifics, we find that the critical value of χ_1^2 (say χ_{1*}^2) is 3.841 for $\alpha = 5$ percent. Since the observed value $\chi_1^2 = 4.00 > \chi_{1*}^2 = 3.841$, we reject H_0 and accept H_A; that is, these observations provide sufficient evidence to say that the coin is biased.

This χ^2 procedure and its result ought to be completely equivalent to the test we had earlier performed by use of Z. It happens that the equivalence is mathematically precise because the probability pattern of Z^2 can be proved to be exactly the pattern of χ_1^2 (χ^2 with 1 d.f.). Thus, the critical values match:

$$z_* = 1.960, \qquad z_*^2 = (1.960)^2 = 3.842; \qquad \chi_{1*}^2 = 3.841,$$

(More nearly exactly, $z_* = 1.959964$, $z_*^2 = 3.841459$), and our calculated statistic values match:

$$z = 2.00; \qquad \chi_1^2 = 4.00 = z^2$$

7.4 1×c TABLE

The extension of the binomial or 1×2 table to the more general $1 \times c$ table is straightforward. Let us illustrate this extension with the following problem:

A manufacturing plant runs 24 hours a day by utilizing three shifts. Each shift has the same number of employees working. The medical doctor for the plant informs the plant manager that he is concerned about the number of minor accidents occurring in the plant. He states that he believes the major problem to be poor safety practice on some of the shifts. After some discussion he furnishes the following data on minor accidents by shift for the period July 1 through December 31, 1972 (Table 7.4.1).

TABLE 7.4.1

	Shift 1	Shift 2	Shift 3	Total
Number of minor accidents	25	30	41	96

If one hypothesizes that safety practices, susceptibility to minor accidents, and other conditions are the same in all shift operations, is there any evidence in these data to refute this hypothesis?

On the basis of this hypothesis, the probability of a minor accident occurring on shift 1 = probability of a minor accident occurring on shift 2 = probability of a minor accident occurring on shift 3, and then the likelihood of a minor accident's being in a particular shift is the same for all three shifts. That is, we have hypotheses as follows:

$$H_0: \; p_1 = p_2 = p_3 = p = \frac{1}{3}$$

versus

H_A: at least two of the subscripted probabilities are unequal.

Let us use $\alpha = 5$ percent for testing this hypothesis. Under H_0, the expected frequencies are calculated and recorded in the following table:

	Shift 1	Shift 2	Shift 3	Total
Observed = o_i	25	30	41	96
Expected = e_i[a]	32	32	32	96

[a] $e_i = p_{oi} \times n = (1/3) \times 96 = 32, \; i = 1, 2, 3.$

Before calculating χ_1^2, let's determine the number of degrees of freedom l. The assumed underlying model has the parameters p_1, p_2, and p_3, subject to the restriction $p_1 + p_2 + p_3 = 1$ (a minor accident that occurs *has* to occur on one of the three shifts). Thus the number of *independent* parameters in the model is 2. Under the null hypothesis H_0, all p_is are equal, and hence have to be 1/3 each. Hence the number of parameters in the assumed underlying model under this specific H_0 is zero. And so $l = 2 - 0 = 2$, telling us that we are dealing with χ_2^2: χ^2 with 2 d.f.

The critical value of χ^2 that has 2 d.f. and cuts off the upper 5 percent of the distribution is $\chi_{2*}^2 = 5.991$.

Calculations now give:

$$\chi_2^2 = \frac{(25-32)^2}{32} + \frac{(30-32)^2}{32} + \frac{(41-32)^2}{32}$$

$$= \frac{49}{32} + \frac{4}{32} + \frac{81}{32}$$

$$= 4.19$$

In conclusion, since $\chi_2^2 = 4.19 < \chi_{2*}^2 = 5.991$, we cannot reject H_0. Thus we say that using these 6 months' data, the three shifts are not significantly different at the 5 percent level. More evidence will be needed if we stick to our 5 percent significance level position and persist in a search for distinctions among the shifts.

This $1 \times c$ categorical procedure is very general: one can hypothesize anything he has the urge to concerning p_1, p_2, p_3, ..., p_c. For example in the above situation involving minor accidents on three different shifts, one could, if he liked, hypothesize that a minor accident is equally likely to be on shift 1 or shift 2 but twice as likely to be on shift 3. Here we would be hypothesizing the p_is to be p_1, p_1, and $2p_1$, and since these must add up to 1, we would have $1 = p_1 + p_1 + 2p_1 = 4p_1$, giving $p_1 = 1/4$, so that our null hypothesis is

$$H_0: \; p_1 = 1/4, \qquad p_2 = 1/4, \qquad p_3 = 1/2$$

versus

$$H_A: \text{not all of these equalities hold.}$$

The test statistic χ^2 will again be χ_2^2: χ^2 with 2 d.f. (the general $1 \times c$ case has $c - 1$ d.f.), and the calculations would be as follows:

	Shift 1	Shift 2	Shift 3	Total
o_i	25	30	41	96
e_i	24	24	48	96

$$\chi_2^2 = \frac{(25-24)^2}{24} + \frac{(30-24)^2}{24} + \frac{(41-48)^2}{48}$$

$$= \frac{1}{24} + \frac{36}{24} + \frac{49}{48}$$

$$= 2.56$$

Since $2.56 < \chi_{2*}^2 = 5.991$, we have to refrain from rejecting H_0 and say that the observed distribution of accidents among the three shifts is not significantly different from that which is hypothesized.

(Note: Having nonrejected two different hypotheses, we can see an argument for avoiding the use of "accept" as the alternative to "reject." Accepting two different hypotheses seems a bit much. But we must remember that a test of hypothesis treats "reject H_0" as meaning "data are *not* probabilistically *consistent* with H_0," and "accept H_0" as meaning "data are *not* probabilistically *inconsistent* with H_0." There can be many hypotheses falling into either category. A statistical test is a valid check of *one* hypothesis with *one* set of data.

Greater precision, greater assurance of correct decision, and greater knowledge about the truth—all require more samples, larger samples, and continuing experimentation. *That* is completely consistent with what scientists have known from earliest times. The contribution of statistical analysis is an orderly procedure, with probability statements giving numerical measure to credibility.)

7.5 2×2 CONTINGENCY TABLE

The next generalization to be considered is the situation where an individual observation is characterized in two ways, each of which is a binomial kind of classification. Consider the following example:

A sample of $n = 78$ housewives was chosen at random and each was asked the following two questions:

1. What brand of soap are you using in your bathroom at the present time?
2. What product do you use for laundering your regular clothes?

Based on a knowledge of the brands made by Procter and Gamble, data from the housewives' responses were tabulated in Table 7.5.1.

TABLE 7.5.1

		Question 1 (Soap)		
		Procter and Gamble Brands	Others	Totals
Question 2 Laundry Product	Procter and Gamble Brands	42	15	57
	Others	10	11	21
	Totals	52	26	78

The question of interest in this problem is whether or not the choice of a bathroom-soap product is independent of the choice of a laundry product.

The null hypothesis H_0 is a statement that the choices *are* independent (i.e., there is *no* association between soap choice and laundry-product choice). The rejection of this hypothesis would be of considerable importance to Procter and Gamble: it would indicate that the Procter and Gamble label in the area of soap products covering both bathroom bars and laundry soaps is important in maintaining market position.

This hypothesis must be written in probability terms as follows:

Let $p_{i.}$ = probability of using laundry products in the ith category:
  $p_{1.}$ = probability of using a Procter and Gamble brand of laundry product,
  $p_{2.}$ = probability of using another brand of laundry product;
 $p_{.j}$ = probability of using bathroom bar soap in the jth category:
  $p_{.1}$ = probability of using a Procter and Gamble bar soap,
  $p_{.2}$ = probability of using another brand of bar soap;
 p_{ij} = probability of using both the ith category laundry product and the jth category bar soap product.
(Note the use of i to index *row* categories in the data table, j to index *column* categories, i number first and j number second. Thus p_{12} = probability of being in the first-row category *and* second-column category. This is standard usage in mathematics generally.)

Using our probability definition of independent events (Chapter 4, Definition 4.4.1), we can state our hypothesis in these terms:

$$H_0: \ p_{ij} = p_{i.} \cdot p_{.j} \qquad \text{for all } i, j \ (i = 1, 2, j = 1, 2)$$

versus

$$H_A: \text{ at least one of the stated equalities does not hold.}$$

We will use the $\alpha = 5$ percent significance level for testing the hypothesis of independence in our example. Since all of these probabilities are unknown and we are not making hypotheses about their specific values, we will estimate them from the data as follows:

First let us repeat the data table with some useful notation added (Table 7.5.1.a).

TABLE 7.5.1.a

		Bar Soap		
		Procter and Gamble Brands	Other Brands	Totals
Laundry Products	Procter and Gamble Brands	$o_{11} = 42$	$o_{12} = 15$	$n_{1.} = 57$
	Other Brands	$o_{21} = 10$	$o_{22} = 11$	$n_{2.} = 21$
	Totals	$n_{.1} = 52$	$n_{.2} = 26$	$n = 78$

Our best estimate of the probability of using a Procter and Gamble laundry brand is:

$$\hat{p}_{1.} = \frac{n_{1.}}{n} = \frac{57}{78}$$

Our best estimate of the probability of using another laundry brand is:

$$\hat{p}_{2.} = \frac{n_{2.}}{n} = \frac{21}{78}$$

$$\left[\text{Notice that } \hat{p}_{1.} + \hat{p}_{2.} = \frac{n_{1.}}{n} + \frac{n_{2.}}{n} = \frac{57}{78} + \frac{21}{78} = 1 \right]$$

Our best estimate of the probability of using a Procter and Gamble bathroom-soap bar is:

$$\hat{p}_{.1} = \frac{n_{.1}}{n} = \frac{52}{78}$$

Our best estimate of the probability of using another bathroom-soap bar is:

$$\hat{p}_{.2} = \frac{n_{.2}}{n} = \frac{26}{78}$$

[Again $\hat{p}_{.1} + \hat{p}_{.2} = 1$]

Based on the null hypothesis H_0: $p_{ij} = p_{i.} \cdot p_{.j}$, we argue $\hat{p}_{ij} = \hat{p}_{i.} \cdot \hat{p}_{.j}$ and then calculate the *expected number* of observations in each of the cells of the table in the following way:

$$e_{ij} = n \times \hat{p}_{i.} \times \hat{p}_{.j},$$

where e_{ij} = expected number of observations in the (i, j)th cell,
n = total sample size.

Thus we obtain:

$$e_{11} = 78 \times \hat{p}_{1.} \times \hat{p}_{.1}$$

$$= 78 \times \frac{57}{78} \times \frac{52}{78} = \frac{(57)(52)}{78} = \frac{2964}{78}$$

$$= 38,$$

$$e_{12} = 78 \times \hat{p}_{1.} \times \hat{p}_{.2}$$

$$= 78 \times \frac{57}{78} \times \frac{26}{78} = \frac{(57)(26)}{78} = \frac{1482}{78}$$

$$= 19,$$

$$e_{21} = 78 \times \hat{p}_{2.} \times \hat{p}_{.1}$$

$$= 78 \times \frac{21}{78} \times \frac{52}{78} = \frac{(21)(52)}{78} = \frac{1092}{78}$$

$$= 14,$$

$$e_{22} = 78 \times \hat{p}_{2.} \times \hat{p}_{.2}$$

$$= 78 \times \frac{21}{78} \times \frac{26}{78} = \frac{546}{78}$$

$$= 7.$$

You may have noticed that each of the above calculations boiled down to row total multiplied by column total divided by grand total. This is no mere coincidence. In every $r \times c$ contingency table the expected numbers are given by:

$$e_{ij} = \frac{n_{i.} \times n_{.j}}{n} \tag{7.5.1}$$

Placing our e_{ij} in the same table with the data, we now have Table 7.5.1.b:

TABLE 7.5.1.b

		Bar Soap		
		Procter and Gamble Brands	Other Brands	Totals
Laundry Products	Procter and Gamble Brands	$o_{11} = 42$ $e_{11} = 38$	$o_{12} = 15$ $e_{12} = 19$	$n_{1.} = 57$
	Other Brands	$o_{21} = 10$ $e_{21} = 14$	$o_{22} = 11$ $e_{22} = 7$	$n_{2.} = 21$
	Totals	$n_{.1} = 52$	$n_{.2} = 26$	$n = 78$

To determine the number of degrees of freedom l, we count the number of independent parameters in the underlying model and subtract the number of independent parameters left after the null hypothesis is applied. In the model we have:

$$p_{11}, p_{12}, p_{21}, p_{22}, \text{ subject to totaling } 1.$$

Another way of saying this is that *every* housewife in the study must fall into one and only one of the four cells. Thus if we write the table in probability notation we would have:

p_{11}	p_{12}	$p_{1.}$
p_{21}	p_{22}	$p_{2.}$
$p_{.1}$	$p_{.2}$	$p = 1$

Thus in the *underlying assumed model* there are four parameters, p_{11}, p_{12}, p_{21}, and p_{22}, *but* they must add up to 1. Thus there are but three *independent* parameters in the underlying model.

Under the *null hypothesis of independence*, H_0 further specifies that $p_{ij} = p_{i.} \times p_{.j}$. As you can see from the above table, $p_{1.} + p_{2.} = 1$ and $p_{.1} + p_{.2} = 1$. Thus of these four, namely $p_{1.}$, $p_{2.}$, $p_{.1}$, $p_{.2}$, only *two* are independent. Then using rule 7.3.2, $l = 3 - 2 = 1$ d.f. Our test statistic is thus χ^2 with 1 d.f.: χ_1^2. The critical value of χ_1^2 when $\alpha = 5$ percent is $\chi_{1*}^2 = 3.841$ (Table A-5).

Using the o_{ij} data and the calculated e_{ij} values in Table 7.5.1.b, we can now calculate χ_1^2.

$$\chi_1^2 = \sum_{\substack{\text{all four} \\ \text{cells}}} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(42 - 38)^2}{38} + \frac{(15 - 19)^2}{19} + \frac{(10 - 14)^2}{14} + \frac{(11 - 7)^2}{7}$$

$$= \frac{16}{38} + \frac{16}{19} + \frac{16}{14} + \frac{16}{7}$$

$$\chi_1^2 = 4.69$$

In conclusion, since the calculated value of $\chi_1^2 = 4.69$ exceeds the critical value of $\chi_{1*}^2 = 3.841$, we reject H_0 and conclude that there is sufficient evidence in these data to say that the choice of bar soap is not independent of the choice of laundry products among the women whom these panel women represent. We say that the observed *association* between selection of bar soap and selection of laundry products is significant at the 5 percent level.

7.6 THE *r*×*c* CONTINGENCY TABLE

The next generalization to be considered is the case where an individual observation is characterized in two ways, each with more than two levels. We think of r categories for the row characteristic and c categories for the column characteristic. For example, a family could be classified by the number of children in the family and by the family income. An example of this is shown in Table 7.6.1.

This type of table is called an $r \times c$ contingency table, in this case a 3×4 table—that is, three rows and four columns, giving a total of $r \times c = 12$ categories.

(The student should note that in this example the columns represent discrete data, while the rows represent an arbitrary division of a continuous variable, income, into three discrete units. This latter procedure is a common one used in many fields, and while it can be useful, any class-intervalizing of data like this reduces the information content of the data. We must weigh the magnitude of this information loss before doing the grouping as illustrated in Table 7.6.1.)

TABLE 7.6.1

Family Income \ Number of Children	0	1	2	Over 2	Totals
Under $2000	$o_{11} = 15$ $e_{11} = 25.2$	$o_{12} = 27$ $e_{12} = 40.4$	$o_{13} = 50$ $e_{13} = 37.3$	$o_{14} = 43$ $e_{14} = 32.1$	$n_{1.} = 135$
2000–5000	$o_{21} = 25$ $e_{21} = 15.3$	$o_{22} = 37$ $e_{22} = 24.6$	$o_{23} = 12$ $e_{23} = 22.7$	$o_{24} = 8$ $e_{24} = 19.4$	$n_{2.} = 82$
Over 5000	$o_{31} = 8$ $e_{31} = 7.5$	$o_{32} = 13$ $e_{32} = 12.0$	$o_{33} = 9$ $e_{33} = 11.0$	$o_{34} = 10$ $e_{34} = 9.5$	$n_{3.} = 40$
Totals	$n_{.1} = 48$	$n_{.2} = 77$	$n_{.3} = 71$	$n_{.4} = 61$	$n = 257$

The data (the o_{ij}s) in Table 7.6.1 were obtained by choosing $n = 257$ families at random, determining family income and family size for each and filling in the table accordingly. Thus the only marginal total fixed in advance was the grand total, $n = 257$.

In this case, the hypothesis of independence is stated:

H_0: Family size and family income are independent of each other.

In terms of probability, let

$p_{i.}$ = probability of the occurrence of the ith family income,
$p_{.j}$ = probability of the occurrence of the jth family size,
p_{ij} = probability of the simultaneous occurrence of the
ith family income *and* the jth family size.

Then our hypothesis is specifically the following:

$$H_0: \ p_{ij} = p_{i.} \times p_{.j} \quad \text{for all } i, j \ (i = 1, 2, 3, j = 1, 2, 3, 4),$$

versus

$$H_A: \text{at least one of the stated equalities does not hold.}$$

Since we have n families, the *expected* number of families falling in the (i, j)th category or cell of the $r \times c$ table is $np_{ij} = n \times p_{i.} \times p_{.j}$ if the null hypothesis H_0 is true.

Since these probabilities are unknown, estimates of them must be obtained from the data:

$$\hat{p}_{i.} = \frac{\text{Number of families falling in the } i\text{th income class}}{\text{Total number of families in the sample}}$$

$$= \frac{n_{i.}}{n};$$

$$\hat{p}_{.j} = \frac{\text{Number of families falling in the } j\text{th family size class}}{\text{Total number of families in the sample}}$$

$$= \frac{n_{.j}}{n}.$$

The expected number of families having the ith income level and the jth family size is then $n\hat{p}_{ij} = n\hat{p}_{i.}\hat{p}_{.j}$, giving

$$e_{ij} = \frac{n_{i.} \times n_{.j}}{n} \tag{7.6.1}$$

as in (7.5.1). Based on the data given in Table 7.6.1, the expected numbers e_{ij} have been calculated and shown in the same table.

To determine the number of degrees of freedom for χ^2, we count independent parameters as before. In the underlying model we have 11:

$$p_{11}, p_{12}, p_{13}, p_{14}, p_{21}, p_{22}, p_{23}, p_{24}, p_{31}, p_{32}, p_{33}, p_{34}, \textit{subject to totaling 1.}$$

According to what H_0 specifies in the model, the parameters are:

$$p_{1.}, p_{2.}, p_{3.}, \textit{subject to totaling 1};$$

$$p_{.1}, p_{.2}, p_{.3}, p_{.4}, \textit{subject to totaling 1.}$$

These show that the *independent* parameters number $2+3=5$. Thus $l = 11-5=6$, and so our χ^2 is χ^2_6 with 6 d.f.

The argument is easily generalized to any $r \times c$ table associated with the null hypothesis of independence. With r rows and c columns, there are rc cells (categories) altogether, and hence rc p_{ij}s, *subject to* the restriction that they add up to 1. Thus, the underlying model has $(rc-1)$ independent parameters.

According to the null hypothesis H_0, the parameters are as follows:

$p_{1.}, p_{2.}, \ldots, p_{r.}$, subject to totaling 1 ($r-1$ independent parameters);

$p_{.1}, p_{.2}, \ldots, p_{.c}$, subject to totaling 1 ($c-1$ independent parameters).

Hence

$$l = (rc-1)-[(r-1)+(c-1)].$$

A bit of algebraic persistence here brings out a beautiful result:

$$l = rc-1-(r+c-2)$$
$$= rc-c-r+1$$
$$= c(r-1)-1(r-1)$$
$$= (r-1)(c-1),$$

and so we have the easy general formula:

$$l = (r-1)(c-1). \tag{7.6.2}$$

Notice that our example, having a 3×4 table, gives $l = (3-1)(4-1) = 2 \cdot 3 = 6$, as we found with harder direct work earlier. Notice also that $l = (r-1)(c-1)$ is precisely the number of e_{ij}s which you can individually compute before the marginal totals force the final cell in each row and each column to come out just right for the sum.

Returning to the test of our example, let us take the level of significance $\alpha = .01$. Then the critical value for the .01 level of significance is $\chi^2_{6*} = 16.812$.

The value of the observed χ^2_6 can now be obtained for testing the hypothesis of the independence of family size and income.

$$\chi^2_6 = \sum_{\substack{\text{all 12} \\ \text{cells}}} \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

A tabulated worksheet will help organize the calculations and speed up the computation by putting the arithmetic on a production-line basis.

Worksheet

o	e	o − e	(o − e)²	(o − e)²/e
15	25.2	−10.2	104.04	4.129
27	40.4	−13.4	179.56	4.445
50	37.3	12.7	161.29	4.324
43	32.1	10.9	118.81	3.701
25	15.3	9.7	94.09	6.150
37	24.6	12.4	153.76	6.250
12	22.7	−10.7	114.49	5.044
8	19.4	−11.4	129.96	6.699
8	7.5	0.5	0.25	.033
13	12.0	1.0	1.00	.083
9	11.0	−2.0	4.00	.364
10	9.5	0.5	0.25	.026

$$\chi_6^2 = 41.248$$

Since $\chi_6^2 = 41.248$ exceeds the critical value 16.812, we reject H_0 and conclude that the observed association between family size and family income is significant at the 1 percent level of significance.

7.7 OTHER USEFUL χ^2 TESTS

There are many other uses for the χ^2 distribution. For example, χ^2 is used for testing the fit of data to a completely theoretical probability distribution, the *goodness of fit* test. It is used to combine the probability levels of several tests of significance. For our purposes, we shall discuss two other interesting uses of the χ^2 statistic: the test of homogeneity and the test for a shift in binomial proportion.

A. Test of Homogeneity (2×2 case)

A test panel of 300 women were asked to try two products, A and B. Fifty percent of the women tried A first and B second, while the other 50 percent tried B first and A second. The data were recorded in Table 7.7.1.

TABLE 7.7.1

	Order A → B	Order B → A	Totals
Preferred A	105	60	165
Preferred B	45	90	135
Totals	150	150	300 = n

Note that one of the margins has fixed totals. (Column totals were fixed by assigning 150 women to Order $A \rightarrow B$ and 150 women to Order $B \rightarrow A$.) When this is true, the hypothesis is framed in the sense of homogeneity. The hypothesis to be tested in this case is: the preference between A and B is the same for those women who tried the products in the order $A \rightarrow B$ as for those who tried the products in the order $B \rightarrow A$.

The procedure of analysis turns out to be the same as the one used to test the independence hypothesis in the 2×2 table discussed previously, although the logic is different.

Here we are checking whether the row-wise category breakdown in column 1 is significantly different from that in column 2. (The query is whether the columns are *homogeneous* as to row breakdown; thus the name of the test.) If the null hypothesis is true, the probability breakdown by row *is* the same for both columns. Thus we would estimate an *overall* probability of being in row 1 (preferred A), and since the overall number of women who preferred A was 165, the estimate would be

$$p_{1.} = \frac{165}{300}$$

Similarly the overall probability of being in row 2 (preferred B) would be estimated by

$$p_{2.} = \frac{135}{300}$$

Then we would *expect* the 150 women in column 1 to be distributed as:

$$e_{11} = 150 \frac{165}{300} = \frac{150 \times 165}{300} = 82.5, \qquad e_{21} = 150 \frac{135}{300} = \frac{150 \times 135}{300} = 67.5,$$

and the 150 women in column 2 to be distributed as:

$$e_{12} = 150 \frac{165}{300} = \frac{150 \times 165}{300} = 82.5, \qquad e_{22} = 150 \frac{135}{300} = \frac{150 \times 135}{300} = 67.5.$$

Note that each e_{ij} comes out to be in accord with our earlier rule (7.6.1): row total times column total divided by grand total.

What about degrees of freedom? The underlying model has the following parameters:

Order $A \rightarrow B$: p_{11}, p_{21}, subject to totaling 1;

Order $B \rightarrow A$: p_{12}, p_{22}, subject to totaling 1.

Thus the model has $1+1=2$ independent parameters. After the null hypothesis H_0 is applied, arguing $p_{11}=p_{12}$ and $p_{21}=p_{22}$, the parameters are:

$$p_{1.}, p_{2.}, \text{ subject to totaling 1.}$$

Hence there is now just *one* independent parameter, and so the rule (7.3.2) for degrees of freedom gives

$$l = 2 - 1 = 1.$$

We note that this is the same as in the 2×2 table testing independence. We wind up the test in the usual manner:

$$\chi_1^2 = \frac{(105-82.5)^2}{82.5} + \frac{(45-67.5)^2}{67.5} + \frac{(60-82.5)^2}{82.5} + \frac{(90-67.5)^2}{67.5}$$

$$= 6.136 + 7.500 + 6.136 + 7.500$$

$$\chi_1^2 = 27.27$$

Since the calculated value of $\chi_1^2 = 27.27$ is greater than the critical value of $\chi_{1*}^2 = 6.635$ for $\alpha = 1$ percent, we conclude that the order of trying the products definitely affects the preference for the products.

B. Test of Homogeneity ($r \times c$ case)

Let us extend the 2×2 case to a more general $r \times c$ situation. Here we consider the question: Is the distribution of a certain characteristic *the same* for several populations? In the previous 2×2 case the question was phrased: "Is the distribution of preference for products A and B in the population of women who try them in the order A then B the same as in the population of women who try them in the order B then A?"

There are many examples of this type of experimental data: a market research company conducts a four-product test in each of five different cities and records the number of people preferring each product; a political pollster selects four different voting districts and gives the sampled people a choice among four different candidates for the presidency in 1976; a pharmaceutical firm tries out three new products for the relief of headaches in people of four different economic levels, each person trying all three products and stating which product he thought best, and so on.

Example 7.7.1

During World War II a survey of men drafted into the Army was conducted on their attitudes toward the draft. It was thought that the amount of prior education before induction into the army would have a strong influence on attitude toward the draft. A random sample of $n_{1.} = 200$ men with less than a ninth-grade education (i.e., having just an elementary education) was drawn, and also random samples of $n_{2.} = 100$ men with a high-school diploma, and $n_{3.} = 50$ men with at least two years of college

education. The data collected are shown as follows

| | | How Do You Feel about the Draft? | | |
		Don't Think It Fair	Think It Fair	Totals
Education	Elementary	$O_{11} = 142$	$O_{12} = 58$	$n_{1.} = 200$
	High School	$O_{21} = 44$	$O_{22} = 56$	$n_{2.} = 100$
	At Least Two Years College	$O_{31} = 8$	$O_{32} = 42$	$n_{3.} = 50$
	Totals	$n_{.1} = 194$	$n_{.2} = 156$	$n = 350$

Is there sufficient evidence in these data to say that the distribution of the attitudes toward the draft differs among the three "level of education" populations?

The null hypothesis is that the distribution is the *same* for all populations. In this example we have taken the populations as given by the *rows* in the table. Under H_0,

$$p_{.j} = \text{probability that an individual from any row population falls in the } j\text{th column.}$$

We estimate $p_{.j}$ by

$$\hat{p}_{.j} = \frac{n_{.j}}{n}, \text{ and}$$

then our *estimate* of the expected number in the (i, j)th cell of the table is

$$e_{ij} = n_{i.} \frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}.$$

Notice that this number is mechanically identical to the e_{ij} obtained in the situation of the χ^2 test of independence: row total times column total divided by grand total.

Let us look at the degrees of freedom generally in an $r \times c$ case like the present one. There are r rows, each representing a particular population. Each row is broken down into c column categories. The parameters in the underlying model are:

Row 1: $p_{11}, p_{12}, \ldots, p_{1c}$, subject to totaling 1,

Row 2: $p_{21}, p_{22}, \ldots, p_{2c}$, subject to totaling 1,

.   .   .   .

.   .   .   .

.   .   .   .

Row r: $p_{r1}, p_{r2}, \ldots, p_{rc}$, subject to totaling 1.

Thus each row contributes $(c-1)$ independent parameters, and, since there are r rows, there are $r(c-1)$ independent parameters altogether. According to the null hypothesis H_0, the probability distribution is the same in every row, and so under H_0 the parameters are:

$$p_{.1}, p_{.2}, \ldots, p_{.c}, \text{ subject to totaling 1.}$$

Hence there are now $(c-1)$ independent parameters, and thus the rule (7.3.2) gives the number of degrees of freedom as:

$$l = r(c-1) - (c-1) = (c-1)(r-1),$$

the same as (7.6.2). If we switch rows and columns in the argument, taking each *column* to represent a population, and the rows to show category breakdown (we had this setup in the 2×2 example of women with product preference), then we get

$$l = c(r-1) - (r-1) = (r-1)(c-1),$$

again the same result as before.

Returning to our data on attitude toward the draft, we calculate the number of degrees of freedom as

$$l = (r-1)(c-1) = (3-1)(2-1) = 2 \cdot 1 = 2,$$

and then proceed with the test:

H_0: Distribution of draft attitudes is the same for all
three educational levels.

If H_0 is true, then

$$\sum_{\substack{\text{all six} \\ \text{cells}}} \frac{(o-e)^2}{e} = \chi^2 \text{ with 2 d.f.}$$

Taking $\alpha = 5$ percent, we set up the decision rule: reject H_0 if $\chi_2^2 \geqslant \chi_{2*}^2 = 5.991$, do not reject H_0 otherwise.

		Draft Attitude		
		Don't Think It Fair	Think It Fair	Totals
Education	Elementary	$o_{11} = 142$ $e_{11} = 110.86$	$o_{12} = 58$ $e_{12} = 89.14$	$n_{1.} = 200$
	High School	$o_{21} = 44$ $e_{21} = 55.43$	$o_{22} = 56$ $e_{22} = 44.57$	$n_{2.} = 100$
	At Least 2 Years of College	$o_{31} = 8$ $e_{31} = 27.71$	$o_{32} = 42$ $e_{32} = 22.29$	$n_{3.} = 50$
	Totals	$n_{.1} = 194$	$n_{.2} = 156$	$n = 350$

$(o-e)$	$(o-e)^2$	$(o-e)^2/e$
31.14	969.6996	8.747
−31.14	969.6996	10.878
−11.43	130.6449	2.357
+11.43	130.6449	2.931
−19.71	388.4841	14.020
+19.71	388.4841	17.429
		56.362

Thus, we have $\chi_2^2 = 56.362$.

The conclusion is to reject H_0. The proportion of men who think the draft was fair is different for the different "educational-level" populations.

C. Test of Shift in Binomial Proportion

A very interesting application of the χ^2 test is the test of whether or not there has been a shift in a binomial proportion. A typical example is the following: One hundred women in a panel test were asked for their preference between two brands of coffee, A and B. For the 2 months following the test, the area from which these women came was bombarded with an advertising campaign on the excellent qualities of brand A coffee. At the end of the 2 months the same 100 women were brought back and given another blind preference test. The results of the two tests were combined and shown in the following table:

		Product Preferred in Second Test		
		Brand A	Brand B	Totals
Product Preferred in First Test	Brand A	$O_{11} = 50$	$O_{12} = 9$	$n_{1.} = 59$
	Brand B	$O_{21} = 21$	$O_{22} = 20$	$n_{2.} = 41$
	Totals	$n_{.1} = 71$	$n_{.2} = 29$	$n = 100$

This table looks like an ordinary 2×2 contingency table but there is a very special peculiarity about it. It is addressed to the query as to whether opinion *before* the advertising campaign is *homogeneous* with opinion *after*. But it cannot allow us a test of homogeneity because we have only one single sample, used both before and after, rather than a sample from the *before* population and an independent sample from the *after* population. The table *can* give us a χ^2 test of independence (one sample categorized in two ways), but here that is of no real interest: we do not need a statistical test to determine that the opinions of a set of people at one time are *associated* with their opinions at another time.

We do want to test the null hypothesis that the probability of a person's preference for brand A *before* the advertising campaign is the same as the probability of a person's preference for brand A *after* the campaign. But with our experimental setup, this means:

$$H_0: p_{1.} = p_{.1},$$

and that is *not* the sort of null hypothesis handled by the standard χ^2 tests on contingency tables.

Theorists have worked out an acceptable test procedure, which has the added advantage of intuitive appeal since it is based on assessing the *switch* votes. If there were *no* change in the overall group preference between brands A and B, we would expect all of those whose preferences switched from brand A to B to be counterbalanced by those whose preferences switched from brand B to A. Thus we consider only the 30 people whose preferences switched. Under H_0, we would expect these to be split 50/50 between the two kinds of switching, and so take $e_{12} = 15$, $e_{21} = 15$. The resulting χ^2 has 1 d.f.

$$\chi_1^2 = \frac{(9-15)^2}{15} + \frac{(21-15)^2}{15}$$

$$= \frac{36}{15} + \frac{36}{15} = 4.80.$$

This result is significant at the 5 percent level since $\chi_1^2 = 4.80$ is greater than the critical value $\chi_{1*}^2 = 3.841$. Thus at that level of significance we reject H_0 and state that the observed change in preference has been significant. This is a good indication, then, that the advertising campaign was indeed successful.

7.8 EXERCISES

7.8.1 It has been conjectured that the "serious" injury that occurs in football competition occurs in the latter part of any specific game. Some data were collected and are shown in the following table:

Time Since Start of Game (minutes)	Number of Serious Injuries
0–21	57
22–42	88

Do you agree that there is a difference between the proportions of injuries in the two halves? Use $\alpha = .05$.

7.8.2   Suppose the table in Exercise 7.8.1 were as follows:

Time Since Start of Game (minutes)	Number of Serious Injuries
0–21	70
22–42	75

At the same level of significance, what would be your conclusion now? If the two tables were based on national experience for the last 2 years, what would be your attitude toward the conjecture stated in Exercise 7.8.1?

7.8.3   A group of individuals, selected at random, are asked a question concerning acceptability of one of a company's products. After an extensive advertising campaign, these same people are asked the question again. The data are recorded as follows:

	Before Campaign		
	Yes	No	Total
After Campaign			
Yes	42	28	70
No	20	30	50
Total	62	58	120

a. Can these data be analyzed as an ordinary contingency table? If so, what hypothesis would be tested?
b. Test the hypothesis of no change in opinion. Use $\alpha = .05$.

7.8.4   In a survey of 300 people in one city, 144 people preferred brand A soap to all others; and in a sample of 600 people in another city, 312 people preferred the same product A. Is this sufficient evidence to reject the hypothesis of equal preference for brand A in both cities? Analyze this using the: (a) binomial and (b) χ^2 distribution. [Use $\alpha = .05$.]

7.8.5   Evidence has demonstrated that among working women, 56 percent of those who stand while working have varicose veins. In a sample of 400 women who sit or walk while at work, 120 of them were found to have varicose veins. With an α risk of 5 percent, is this sufficient evidence to indicate that working conditions may affect the occurrence of varicose veins?

7.8.6   In repeated surveys in the United States, 70 percent of high-school students answer "a lot" to the question "How much does it matter to a person like yourself whether you're healthy or not?" In a special survey among 40 ghetto high-school students, 80 percent stated that their health meant "a lot" to them. Is this sufficient evidence at a risk of $\alpha = 5$ percent to indicate that ghetto children are more concerned about their health than the average respondent?

7.8.7 Injuries classified as "serious" in a particular article [Kraus, J. F., and Gullen, W. H., An epidemiological investigation of predictor variables associated with intramural touch football injuries, *Am. J. Public Health 59* (12), 2144 (1969)], were conjectured to have occurred more frequently in the later stages of a game. Table 10 in that article shows the following data: *

Time Since Start of Game (minutes)	Number of Injuries
0–7	23
8–14	17
15–21	17
22–28	24
29–35	34
36–42	30
Total	145

Using an α risk of 5 percent, do you agree with this conjecture?

7.8.8 In Example 6.10.2 we tested the hypothesis that 25 percent of the student population from which we drew the Mecca Community College sample are unfavorable to legalizing marijuana, against the alternative that the percentage is something other than 25 percent. Test the same hypothesis against the same alternative, with the same level of significance (5 percent), using the χ^2 test.

7.8.9 Three hundred and eighty-eight secondary students turned in a self-administered health questionnaire. Physicians' ratings of the same 388 students were used as a criterion to validate their responses. The total number of "yes" answers in the questionnaire was recorded; each "yes" indicates the presence of a specific health problem. The data were recorded in the following table:

Number of Questionnaire "Yes" Responses	Physicians' Classification			Total
	No Problem	Problem No Priority	Problem With Priority	
0–5	33	28	47	108
6–12	40	32	69	141
13–69	20	42	77	139
Total	93	102	193	388

Using a χ^2 test with $\alpha = 5$ percent, determine whether the students' ideas of their health status are significantly related to the physicians' rating of their health status.

*Copyright © 1969 by the American Public Health Association, Inc. Reprinted by permission of the author and publisher.

7.8.10   The data from a study of the incidence of pneumoconiosis among Appalachian bituminous coal miners show a relation between history of dust on the lung and degree of pneumoconiosis. Among those working miners in the study the following table was constructed.

Dust on Lung	Degree of Pneumoconiosis				Totals
	None	Suspect	Simple	Complicated	
Positive history	106	14	39	38	197
No history	2031	121	133	38	2323
Totals	2137	135	172	76	2520

Determine whether there is significant association between pneumoconiosis and dust on the lung. What happens if you collapse the pneumoconiosis scale to the same type of scale as dust on lung? Discuss how you would suggest doing this. Using your reduced table, is there significant association between dust on lung and pneumoconiosis? (Use $\alpha = 5$ percent.)

7.8.11   Another table in the study described above shows a relation between roentgenographic findings and principal occupation. The following extracted data will be used:

Occupation	Degree of Pneumoconiosis			Total
	None	Suspect	Definite	
Working miner	2143	134	248	2525
Nonworking miner	872	79	213	1164
Totals	3015	213	461	3689

Working Miner	Degree of Pneumoconiosis			Total
	None	Suspect	Definite	
Surface	375	8	10	393
Underground	1768	126	238	2132
Totals	2143	134	248	2525

a. Using $\alpha = 5$ percent, determine whether the degree of pneumoconiosis is significantly different for working and nonworking miners.
b. Perform the same sort of test for the surface and underground working miner.
c. Draw conclusions about working and nonworking miners and the degree of pneumoconiosis.

8

Predicting
With
Confidence

8

8.1 INTRODUCTION

In this chapter we shall discuss a different kind of estimation problem. Suppose we were asked to estimate how much money a typical family of four saved each year. In thinking about this problem, it would occur to us that the amount of money saved by the family would certainly depend on how much money the family earned. If we knew the family earnings we could probably predict the savings much better. Let's consider our Mecca Community College data— one might conjecture that a student's G.P.A. might be affected by the time the student takes in commuting to school.

Thus these kinds of problem require the knowledge of two pieces of data for every observation. In the family savings example, for each family we need the amount of earnings, call it x, and the amount of savings, call it y. In the Mecca College example, for each student we need the commuting distance, x, and the G.P.A., y.

In general, then, we consider a slightly more complex situation. The random variable, Y, is thought to be better explained by considering its relationship to another measure, call it X. Our procedure will be to build a model relating Y to X (sometimes called a *fixed* variable). In this chapter we shall deal only with the simplest relationship between Y and X, namely a straight line. We shall proceed slowly in building this relationship or model. Our method is to use a real example with some real data.

8.2 AN EXAMPLE

We all know that the older an automobile is the more it costs to keep it in good operating condition. It has been hypothesized that the maintenance cost of an automobile rises at a steady rate through the first 3 years of operation. In order to ascertain the correctness of this hypothesis, the following data were collected from six car owners on the cost of maintaining their automobiles:

Age of Car (in months)	Maintenance Cost ($ per 6 months)
x	y
6	50
12	75
18	100
24	175
30	200
36	300

8.3 FITTING AN EQUATION TO THE DATA

We shall proceed step by step toward building a good predictive model, that is, an equation or formula for predicting how much the semiannual maintenance cost will be at any age of the car.

STEP 1. First let's ignore how old the cars are, and just determine some characteristics of the data we have on cars. In other words, consider that we have a random sample of $n = 6$ semiannual car maintenance costs. Let's use what we've learned to date and characterize the sample data by calculating the mean, $\bar{y}$.

$$\bar{y} = \frac{\Sigma y}{n} = \frac{\$900}{6} = \$150$$

Thus our first predictive model for the semiannual maintenance cost of automobiles varying in age from 6 to 36 months is $150.

In modeling terms, we have adopted a very simple model for which the observed data are: $y_i = \mu + e_i$, $i = 1, 2, \ldots, 6$, the e_i values being random deviations around the mean μ.

We estimate μ, the population mean semiannual cost, with $\bar{y}$, and our first prediction model is

$$\hat{y} = \bar{y} \qquad (8.3.1)$$

that is,

$$\hat{y} = \$150$$

where the circumflex ($\hat{\ }$) or "hat" indicates that y is to be predicted by $\hat{y}$.

STEP 2. How good is this first predictive model (8.3.1)? Once we've established this simple model, we should determine how good a fit to reality it gives. In other words, how much do the actual observed costs differ from $\hat{y}_i$ if every $\hat{y}_i$ is taken as $150, in accordance with this model? And what is the overall measure of the discrepancy?

One easy way to decide this is to calculate how much of the *crude* variation in y (variation of y from zero) is explained by the model (8.3.1). We do this by drawing vertical lines from the x axis to the points (x_i, y_i), squaring the distances given by these lines, and tallying these squared distances (Figure 8.3.1).
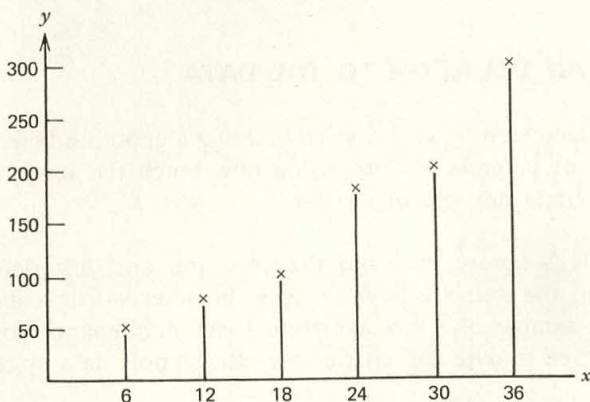
FIGURE 8.3.1

The result of the operation on the observed data points is:

$$\sum y_i^2 = (50)^2 + (75)^2 + (100)^2 + (175)^2 + (200)^2 + (300)^2$$
$$= 178{,}750.$$

Thus we say that the amount of crude variation in costs for our data set is measured as 178,750.

If we replace these individual costs by saying "the average semiannual maintenance cost for these cars is $150," then we're using the predictive model $\hat{y}_i = 150$ for all i. Graphically, we're replacing Figure 8.3.1 with Figure 8.3.2.
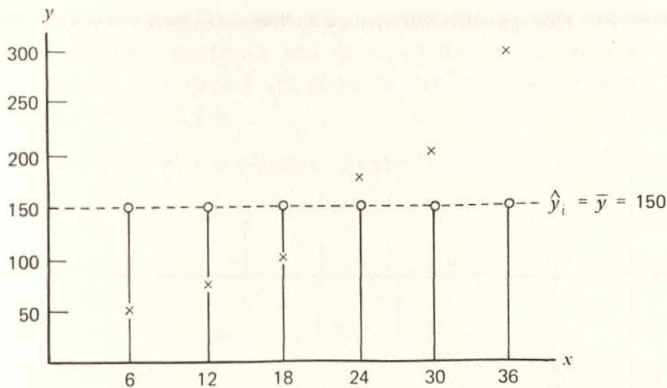
FIGURE 8.3.2

The amount of crude variation explained by this predictive model ($\hat{y}_i = 150$) is:

$$(150)^2 + (150)^2 + (150)^2 + (150)^2 + (150)^2 + (150)^2 = 6(150)^2$$
$$= n(\bar{y})^2$$
$$= 135,000$$

In general, this kind of sum ($n\bar{y}^2$) is called *the sum of squares due to the mean.* We can summarize our modeling status as follows:

a. Using the data, we've established a predictive model, $\hat{y} = 150$, the mean of the observations.

b. Of the total crude variation of 178,750, this model explains 135,000. Thus the model explains

$$\frac{135,000}{178,750} \times 100 = 75.5 \text{ percent of the crude variation}$$
$$\text{in our data.}$$

c. There remain $178,750 - 135,000 = 43,750$ squared units to explain.

Graphically, this can be shown as follows (Figure 8.3.3):
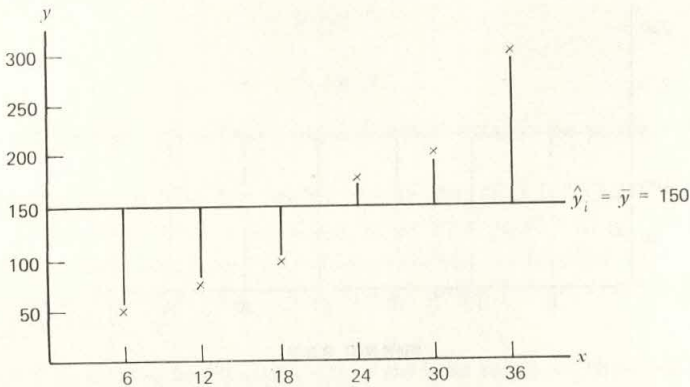
FIGURE 8.3.3

d. The deviations of the observations y_i from the first predictive model $\hat{y}_i = \bar{y} = 150$ behave as follows:

	(1) x_i	(2) y_i	(3) $\hat{y}_i = \bar{y}$	(4) $y_i - \bar{y}$	(5) $(y_i - \bar{y})^2$
	6	50	150	−100	10,000
	12	75	150	−75	5,625
	18	100	150	−50	2,500
	24	175	150	+25	625
	30	200	150	+50	2,500
	36	300	150	+150	22,500
Totals	126	900	900	0	43,750

STEP 3. Now, we notice that there is a pattern to the remaining unexplained deviations (column 4). They are closely related to the values of x in column 1. As x increases, the deviations increase steadily from −100 to +150. Since $\bar{y} = 150$ is a constant, the pattern we see in column 4 exists also in the original data. As a matter of fact, a practical look at Figure 8.3.1 suggests that a *straight line* would fit the data pretty well.

We recall that the equation of a straight line has the form

$$y = \beta_0 + \beta x, \qquad (8.3.2)$$

where β_0 and β are numerical constants which identify the specific line shown in Figure 8.3.4, where β_0 is the value for y when $x = 0$.

From Figure 8.3.4 we can see why β_0 is called the y *intercept* of the line: β_0 is the vertical distance from the x axis at which the line crosses (intercepts) the y axis. The constant β is called the *slope* of the line since it gives the rise (or fall) in y per unit change in x.

In our case we have two problems. In the first place, we don't know what
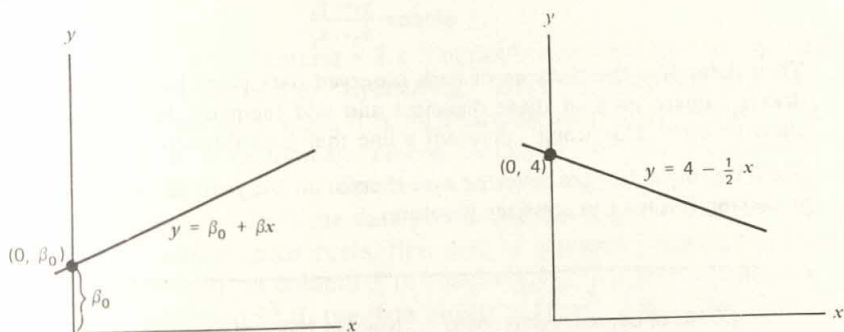
FIGURE 8.3.4

values of β_0 and β are being used by Nature in the process we are studying, and in the second place, Nature uses that line only as an average and deals us observations at variable locations off the line. So we have to use our observed data and *estimate* the line—that is, estimate β and β_0. Let's plot the data (Figure 8.3.5).

We've already decided to use $\hat{y}_i = \bar{y}$ as a first step. Thus let's make sure that the new line goes through the point $(\bar{x}, \bar{y})$. In our data this point is (21 months, \$150). It is shown in Figure 8.3.5 as an open square with a dot in the center.

There are many possible straight lines that can be drawn through the point $(\bar{x} = 21, \bar{y} = 150)$. We would like to choose the line which comes "closest" to all the points.
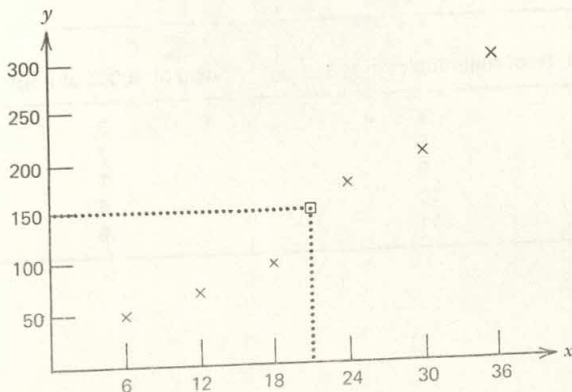
FIGURE 8.3.5

EXERCISES

8.3.1 Draw a line by eye that fits the car maintenance data as best you can. Make sure it goes through the point ($\bar{x} = 21$, $\bar{y} = 150$). Determine the slope of the line you've drawn using any two points (x_1, y_1) and (x_2, y_2) by the formula:

$$\text{slope} = \frac{y_2 - y_1}{x_2 - x_1}$$

Then determine the distance of each observed data point from the line you've drawn. Square each of these distances and add them up. How close did you come to zero? This would represent a line that goes through every data point.

8.3.2 The following data were collected by a chemist on the yield of a production plant at varying levels of process temperature.

x (in 10s of degrees Fahrenheit)	y (yield in 1000s of pounds)
7	1
8	3
9	5
10	5
11	6

a. Putting temperature on the horizontal axis and yield on the vertical axis, plot the five (x, y) points.
b. Calculate $\bar{x}$ and $\bar{y}$. Draw by eye a line through $(\bar{x}, \bar{y})$ to fit the data.
c. Do you think temperature is a pretty good predictor of yield in this process?

8.3.3 Another chemist in the same plant mentioned in Exercise 8.3.2 gathered data on pressure and yield. His data were as follows:

x (10s of millimeters of mercury)	y (yield in 1000s of pounds)
7	3
8	2
9	1
10	4
11	5

a. Putting pressure on the horizontal axis and yield on the vertical axis, plot the five (x, y) points.
b. Calculate $\bar{x}$ and $\bar{y}$.
c. By eye, draw a best-fitting line to the data through $(\bar{x}, \bar{y})$.
d. Do you think this line is as good as that in Exercise 8.3.2?
e. Which is a better *linear* (straight line) predictor of yield—temperature or pressure?

The line you drew in Exercise 8.3.1 is undoubtedly very close to the best that can be done. Your process of "eyeballing," however, does not allow scientific measurement of just how good it is. In fitting straight lines to data, there *is* a procedure that allows such measurement, and moreover gets the prize for "best fit." We say *best fit* here in the sense of *minimizing* the sum of squared distances that exist between the data points and the fitted line. In our example on automobile-maintenance costs, that sum of squared distances, remember, was 43,750 [$\sum (y - \bar{y})^2$ in column 5 of subparagraph (d) above] when we fitted just the mean line $\bar{y} = 150$ to the data points in Figure 8.3.3. The "line of best fit" will give a much smaller sum of squared deviations.

The complete official name of the line to which we are referring is *the line of best fit according to the principle of least squares*. The expression "least squares" is shorthand for "smallest possible sum of squared vertical deviations between the data points and the line." Long a common tool of widespread use among scientists of all kinds, the line is customarily referred to more briefly as the *least-squares line* or the *least-squares line of best fit*.

Working out the mathematical mechanics to produce a formula for the least-squares line is an exercise in algebra or calculus which need not concern us here. The result is easy to state and use. The least-squares line goes through the point of means $(\bar{x}, \bar{y})$ and has slope given by the formula

$$ b = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} $$

Like the sum of squares for y that we use in calculating the sample variance s^2, the above formula can be put into other forms somewhat more convenient for computing. The entire recipe for the *least-squares line* can be stated as follows:

Least-squares line

$$\hat{y} = \bar{y} + b(x - \bar{x}),$$

where   (8.3.3)

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

After the equation for the line has been written in accordance with (8.3.3), it can be put in the form of (8.3.2) by multiplying out and collecting terms:

$$\hat{y} = (\bar{y} - b\bar{x}) + bx$$

Here we see b as our *estimate* of β, and $(\bar{y} - b\bar{x})$ as our *estimate* of β_0.

Let us apply this procedure to our data on automobile-maintenance costs (Table 8.3.1).

TABLE 8.3.1

	x_i	$x_i - \bar{x}$	y_i	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
	6	−15	50	−100	1500	225
	12	−9	75	−75	675	81
	18	−3	100	−50	150	9
	24	+3	175	+25	75	9
	30	+9	200	+50	450	81
	36	+15	300	+150	2250	225
Totals	126	0	900	0	5100	630
Means	21		150			

$$b = \frac{5100}{630} = 8.10$$

This means that the maintenance cost increases approximately $8.10 each month during the first 3 years (36 months) of the automobile's life.

Calculations making use of the alternative computing form of b are as follows (Table 8.3.2). [We carry along y^2 calculations because they will be needed sooner or later.]

TABLE 8.3.2

x	y	x^2	y^2	xy
6	50	36	2,500	300
12	75	144	5,625	900
18	100	324	10,000	1,800
24	175	576	30,625	4,200
30	200	900	40,000	6,000
36	300	1,296	90,000	10,800
Totals 126	900	3,276	178,750	24,000

$$\bar{x} = \frac{126}{6} = 21.0, \qquad \bar{y} = \frac{900}{6} = 150.0,$$

$$\sum (y - \bar{y})^2 = \sum y^2 - n\bar{y}^2 = 178,750 - 6(150.0)^2 = 178,750 - 135,000 = 43,750$$

$$\sum (x - \bar{x})^2 = \sum x^2 - n\bar{x}^2 = 3276 - 6(21.0)^2 = 3276 - 2646 = 630$$

$$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - n\bar{x}\bar{y} = 24,000 - 6(21)(150) = 24,000 - 18,900$$

$$= 5100$$

$$b = \frac{5100}{630} = 8.10$$

The equation of the least-squares line is

$$\hat{y} = 150 + 8.10(x - 21)$$

STEP 4. How good is this predictive model?

We recall that we had 43,750 squared units of variation left over (i.e., unexplained) after fitting the model $y_i = \mu + e_i$ and obtaining the fit (8.3.1) $\hat{y}_i = 150$. The question to be asked is "how much of the 43,750 squared units has been explained by model (8.3.2) for which the fit by (8.3.3) is $y_i = 150 + 8.10(x_i - 21)$?" Let us calculate the fitted value $\hat{y}_i$ for each x_i in the data set. The details are worked out in Table 8.3.3.

TABLE 8.3.3

x_i	y_i	$\hat{y}_i = 150 + 8.10(x_i - 21)$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
6	50	28.5	21.5	462.25
12	75	77.1	-2.1	4.41
18	100	125.7	-25.7	660.49
24	175	174.3	0.7	0.49
30	200	222.9	-22.9	524.41
36	300	271.5	28.5	812.25
Totals 126	900	900	0	2,464.30

Thus we see that of the 43,750 squared units, only 2464.30 are left over (unexplained). We can say that fitting model (8.3.2) to the data explains

$$\frac{43,750 - 2464.30}{43,750} \times 100 = \frac{41,285.70}{43,750} \times 100 = 94.4 \text{ percent}$$

of the variation in y remaining after applying $\bar{y}$ to the data (see Figure 8.3.6).
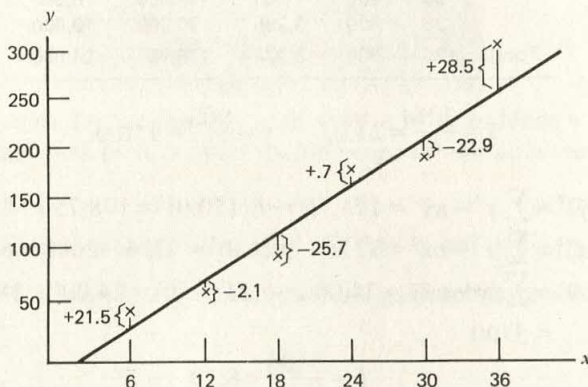
FIGURE 8.3.6

STEP 5. Using the above procedure to fit a line to y data influenced by a characteristic x, we have advanced from the simple model

$$Y_i = \mu + E_i, \qquad i = 1, 2, \ldots, n$$

to the better-fitting model based on (8.3.2):

$$Y_i = \mu + \beta(x_i - \bar{x}) + E_i, \qquad i = 1, 2, \ldots, n, \qquad (8.3.4)$$

or

$$Y_i = \beta_0 + \beta x_i + E_i, \qquad i = 1, 2, \ldots, n, \qquad (8.3.4')$$

where $\beta_0 = \mu - \beta\bar{x}$. In this model the random variation is contributed by the random variables $E_1, E_2, \ldots, E_n$. These are considered to be independent random variables, each having mean zero and variance σ^2. This tells us that $Y_1, Y_2, \ldots, Y_n$ are independent random variables such that

$$\text{mean of } Y_i = \mathscr{E}(Y_i) = \beta_0 + \beta x_i,$$

$$\text{variance of } Y_i = \sigma^2 \text{ for every } i.$$

Our *estimate of the mean* $(\beta_0 + \beta x_i)$ is precisely the least-squares fit (8.3.3):
$$\hat{y}_i = \bar{y} + b(x_i - \bar{x}).$$

The true population *mean* line
$$y = \beta_0 + \beta x$$
is called the *regression line of y on x*. The least-squares line (8.3.3) is then often referred to as the *estimated regression line of y on x*. In the sense of these terms, the relationship of y to x is called *linear regression*.

We can now summarize our findings in Table 8.3.4, called an Analysis of Variance (ANOVA) table. In this table s^2 = residual variance is a valid estimate of σ^2, the unknown population variance. The $\sqrt{s^2} = s$ is an estimate of the population standard deviation, σ.

TABLE 8.3.4 Analysis of Variance

Source of Variation	Degrees of Freedom (d.f.)	Sum of Squares (S.S.)	Mean Square (Sum of Squares ÷ d.f.) (M.S.)
Total (crude)	$n = 6$	$\sum y^2 = 178{,}750$	
Sample mean	1	$n\bar{y}^2 = 135{,}000$	
Total (corrected for the mean)	$n - 1 = 5$	$43{,}750$	
Slope (regression)	1	$41{,}285.70$	$41{,}285.70$
Residual	4	$2{,}464.30$	$616.075 = s^2$

In our example,
$$s^2 = 616.075,$$
$$s = 24.82.$$

Notice that the *sum of squares* for residual (2,464.30 in the above example) is exactly the measure of unexplained variation which is worked out in the manner of Table 8.3.3. Thus we have the meaningful formula:

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - 2}. \tag{8.3.5}$$

Also, it can be shown by algebraic manipulations that

$$\begin{bmatrix} \text{S.S. for total} \\ \text{(corrected for the mean)} \end{bmatrix} = \begin{bmatrix} \text{S.S. for} \\ \text{regression} \end{bmatrix} + \begin{bmatrix} \text{S.S. for} \\ \text{residual} \end{bmatrix}$$

is expressible as

$$\sum (y_i - \bar{y})^2 = b \sum (x_i - \bar{x})(y_i - \bar{y}) + \sum_i (y_i - \hat{y}_i)^2. \qquad (8.3.6)$$

It is often more convenient computationally to use (8.3.6) to compute first the sum of squares for regression and then get by subtraction the sum of squares for residual.

8.3.4   a. Using the format in Table 8.3.1, determine b, the slope of the best-fitting straight line to the data of Exercise 8.3.2, repeated here for your convenience.

x	y
7	1
8	3
9	5
10	5
11	6

   b. Write the equation of best fit as

$$\hat{y} = \bar{y} + b(x - \bar{x})$$

   c. When $x = 0$, determine the value of $\hat{y}$. This is called the *intercept.*

8.3.5   Using the following data, determine the slope b and the intercept of the best-fitting straight line for predicting yield as a function of pressure. (Use the format in Table 8.3.2.)

x (10s of millimeters of mercury)	y (Yield in 1000s of pounds)
7	3
8	2
9	1
10	4
11	5

8.3.6 Using the results of Exercise 8.3.4 above:

    a. Calculate $\hat{y}$ (the predicted value of y) for each x value in the data.
    b. Calculate $y_i - \hat{y}_i$ for each of the points in the data set.
    c. Do you see any patterns in these residuals that would make you reject the straight line as good prediction equation?

8.3.7 Using the data and results of Exercise 8.3.5 above:

    a. Calculate the predicted value of y for each value of x in the data set.
    b. Calculate the residuals $y_i - \hat{y}_i$ for each data point.
    c. Do you see any patterns in these residuals? If so, what do you think of the straight line as a predictor?

8.3.8 The following data table was given for fitting a straight line to a set of data. The best-fitting straight line was $\hat{y} = 10 - .82(x - 7)$.

x_i	y_i	$y_i - \bar{y}$	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	y_i^2	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	14	4	14.92	-0.92	0.8464	196	-6	36	-24
3	15	5	13.28	1.72	2.9584	225	-4	16	-20
5	11	1	11.64	-0.64	0.4096	121	-2	4	-2
7	10	0	10.00	0	0	100	0	0	0
9	9	-1	8.36	0.64	0.4096	81	+2	4	-2
11	5	-5	6.72	-1.72	2.9584	25	+4	16	-20
13	6	-4	5.08	0.92	0.8464	36	+6	36	-24
Totals 49	70	0	70.00	0.00	8.4288	784	0	112	-92

    a. Using the format shown in Table 8.3.4, construct the analysis of variance summary for this model.
    b. Calculate s^2, the estimate of the unknown variance σ^2.
    c. Calculate s, the standard deviation.
    d. Using the formula shown in Step 4, how much of the variation in y is explained by fitting the line's slope to the data? The formula is written as follows:

$$\frac{\text{Sum of squares due to slope}}{\text{Total (corrected for mean) sum of squares}} \times 100 \text{ percent}$$

    e. Do you think this straight line is a good prediction equation for these data?

The regression model (8.3.4) is a better model than the simple $Y_i = \mu + E_i$ if and only if β is not zero. Our estimate of β is the value b computed as the slope of the least-squares line. We shall want to test whether this value is "significantly different" from zero. The logic of Chapter 6 will apply; we shall need a few new details.

Should we decide that b is significantly different from zero, then we will want to use the least-squares line (8.3.3) as a *predictor* for Y:

$$\hat{y} = \bar{y} + b(x - \bar{x}).$$

The logic of Chapter 5 will guide us in formulating confidence intervals for estimation. We shall be careful about necessary details.

The theory that we want to apply requires that the data come from a population having a *normal* distribution, and we now make that assumption. Specifically, we assume that the observations y_i are a random sample of observations from normal distributions whose true means fall on the true line: mean of $Y = \beta_0 + \beta x$, and for which the standard deviation is σ at every value of x. This can be shown graphically as follows:
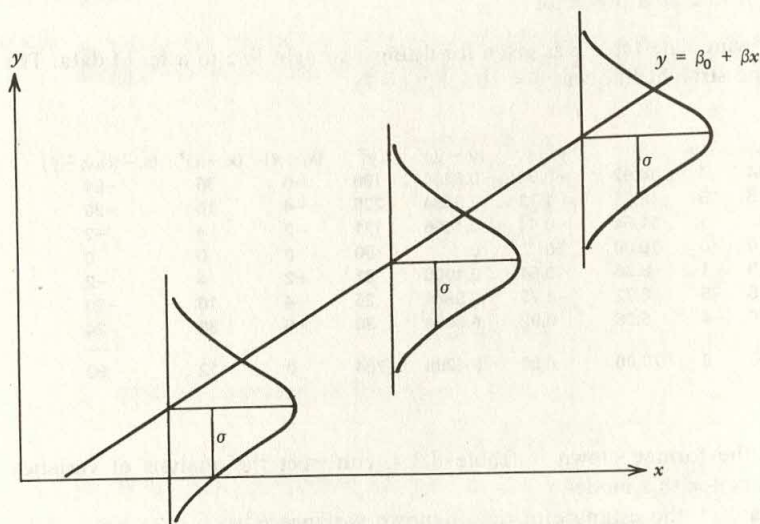
Thus, after fitting our least-squares line to the data, the square root of the residual mean square $= \sqrt{s^2} = \sqrt{616.075} = 24.82$ is an estimate of the σ shown in the above figure.

8.4 TEST OF HYPOTHESIS ABOUT β, THE SLOPE OF THE POPULATION REGRESSION LINE

We can now state our hypothesis about the slope β of the true underlying line:

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_A : \beta \neq 0$$

With the assumptions we have made about the distribution of the Y_is, we can assert that the estimator **b** has a normal distribution in which the mean is β and the standard deviation, say σ_b, is

$$\sigma_b = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} \qquad (8.4.1)$$

We estimate σ_b by the *sample standard deviation* of b:

$$s_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}, \qquad (8.4.2)$$

where s is the square root of the residual mean square in the ANOVA table. The theory then states:

$$\frac{b - \beta}{s_b} = \frac{b - \beta}{s/\sqrt{\sum (x_i - \bar{x})^2}} = t \text{ with } n - 2 \text{ d.f.} \qquad (8.4.3)$$

Test of $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$ can now proceed as in Chapter 6. Let us take $\alpha =$ level of significance at 5 percent.

If H_0 is true, then

$$\frac{b - 0}{s/\sqrt{\sum (x - \bar{x})^2}} = t \text{ with } 4 \text{ d.f.}$$

Reject H_0 if $t \leqslant -2.776$ or if $t \geqslant +2.776$; accept H_0 otherwise. From the sample data, giving the results shown in Section 8.3, we have

$$t = \frac{8.10 - 0}{\frac{24.82}{\sqrt{630}}} = \frac{8.10}{\frac{24.82}{25.10}} = \frac{8.10}{0.9888} = 8.19$$

Reject H_0. The observed regression line slope is significantly different from zero, at the 5 percent level of significance.

The conclusion reached here (observed slope significantly different from zero) is often expressed as *the observed regression is significant.*

As in other tests of significance, we can state the descriptive level of significance P. Here it is

$$P = 2P(t \geqslant 8.19 \mid 4 \text{ d.f.}); \qquad .001 < P < .01.$$

8.5 INTERVAL ESTIMATE FOR β

If the slope is not zero, what is it? Let us make a 95 percent confidence statement about the true slope, β.

Now that we've concluded the true slope is not zero, our best point estimate of β is $b = \$8.10/$month. Using procedures we've learned in Chapter 5, let's place confidence limits on β. Recall the general form of the confidence interval when σ^2 is unknown,

$$\text{parameter estimate} - t_* \left\{ \begin{array}{c} \text{standard deviation} \\ \text{of the} \\ \text{parameter estimate} \end{array} \right\} < \begin{array}{c} \text{parameter} \\ \text{being} \\ \text{estimated} \end{array}$$

$$< \begin{array}{c} \text{parameter} \\ \text{estimate} \end{array} + t_* \left\{ \begin{array}{c} \text{standard deviation} \\ \text{of the} \\ \text{parameter estimate} \end{array} \right\}$$

where t_* is the chosen value of t in the t distribution that defines the region of interest.

For our interval estimate of β, the true slope, we have:

$$b - t_* s_b < \beta < b + t_* s_b \qquad (8.5.1)$$

To place 95 percent confidence limits on β, we choose $t_* = 2.776$. (This *must* be the same t used for the critical region in the test of hypothesis at the 5 percent level of significance. This will ensure consistent results.) The 95 percent confidence interval for β is then:

$$8.10 - (2.776)(.9888) < \beta < 8.10 + (2.776)(.9888)$$

$$8.10 - 2.74 < \beta < 8.10 + 2.74$$

$$5.36 < \beta < 10.84 \qquad (8.5.2)$$

Note that this interval does not include zero. If it did, we could *not* have rejected the null hypothesis, namely $H_0 : \beta = 0$.

EXERCISES

8.5.1 Using the data in Exercise 8.3.7 and the calculations obtained:

  a. Use the residuals $y_i - \hat{y}_i$ to obtain an estimate of σ^2. (Hint: use format in Table 8.3.3 for the formula (8.3.5) for s^2.)

  b. Determine how many degrees of freedom there are in estimating σ^2. Explain the reasoning for this answer.

  c. Calculate the estimate of the standard deviation of the slope, s_b, using equation (8.4.2).

  d. Using $\alpha = 5$ percent, test the hypothesis that the true underlying slope β is really zero, that is, $H_0: \beta = 0$ versus the alternative $H_A: \beta \neq 0$. [Hint: use formula (8.4.3).]

  e. State why you used a t test rather than a z test.

8.5.2 In Exercise 8.3.8, the residual standard deviation is $s = 1.30$. Using this figure and additional information obtained in Exercise 8.3.8, place 95 percent confidence limits on the true unknown slope β. Based on this interval estimate, would you reject $H_0: \beta = 0$ and accept $H_A: \beta \neq 0$ with $\alpha = 5$ percent?

  a. What are the degrees of freedom for s?

  b. What value of t_* did you use? How did you find it?

  c. What was the value of $\sum (x_i - \bar{x})^2$? Where is this figure used?

8.6 PREDICTING THE AVERAGE RESPONSE AT A GIVEN VALUE OF x, SAY x_k

If you desire an average response at a particular value of x, say x_k, the point estimate is given by substituting the particular value x_k in the equation of the best-fitting least-squares line:

$$\hat{y}_k = \bar{y} + b(x_k - \bar{x}). \tag{8.6.1}$$

For example, suppose we want to predict the average maintenance cost per month for a car 24 months old. Then in

$$\hat{y} = 150 + 8.10(x - 21)$$

we take $x = x_k = 24$, and obtain

$$\hat{y}_k = 150 + (8.10)(24 - 21)$$

giving the final result

$$\hat{y}_k = \$174.30 \tag{8.6.2}$$

In obtaining this predicted value $\hat{y}_k$, we used the least-squares equation $\hat{y} = \bar{y} + b(x - \bar{x})$. In this equation are two estimates, $\bar{y}$ and b. Each of these estimates was obtained using a sample of six cars between 6 and 36 months of age. One realizes that the particular $\bar{y}$ and b obtained from these six cars are just one set of possible results. If we were to get another sample of cars covering the same age span and assuming the linear relationship relating

maintenance cost and age to be correct, we would obtain another set of estimates $\bar{y}$ and b. Thus we are saying that $\bar{y}$ has a sampling variance and so does b. Using both these estimates in predicting an average result at $x = x_k$ means that the predicted value $\hat{y}_k$ has a variance that includes them both.

We now write down the variance of $\hat{y}_k$ as follows:

$$\text{Var}(\hat{y}_k) = \text{Var}\left[\bar{y} + b(x_k - \bar{x})\right]$$

$$= \text{Var}(\bar{y}) + (x_k - \bar{x})^2 \text{Var}(b)$$

$$= \frac{\sigma^2}{n} + (x_k - \bar{x})^2 \left[\frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\right] \quad (8.6.3)$$

Some of this should look familiar: $\text{Var}(\bar{y}) = \dfrac{\sigma^2}{n}$ comes from as far back as Chapter 5 in our study; the value of $\text{Var}(b)$ appeared recently above, in (8.4.1). Why everything gets put together in the manner shown is another matter—and a matter whose proof goes beyond what is either convenient or instructive in this introduction to statistical inference. So, as you have already done on a number of occasions, take our word for the accuracy of the stated formula.

Since we do not know σ^2, we will substitute our best estimate of σ^2, namely s^2, the residual mean square obtained in the ANOVA table following the fitting of the least-squares line. We will put a hat ($\hat{}$) on the Var to show that we are *estimating* that variance, and set down the practical formula for our use:

$$\widehat{\text{Var}}(\hat{y}_k) = s^2\left[\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x - \bar{x})^2}\right] \quad (8.6.4)$$

From this of course we have the estimated standard deviation (standard error) of $\hat{y}_k$ as the square root:

$$s_{\hat{y}_k} = \text{estimated standard deviation of } \hat{y}_k = s\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x - \bar{x})^2}} \quad (8.6.5)$$

In (8.6.2) we reached the predicted value for the mean maintenance cost of a car 24 months old: $\hat{y}_k = \$174.30$. If we will now go back through our various calculations and take what we need, we can work out the estimated standard deviation of the prediction according to (8.6.5):

$$\text{Estimated standard deviation of } \{\hat{y}_k \mid x = 24\} = 24.82\sqrt{\frac{1}{6} + \frac{(24-21)^2}{630}}$$

$$= 24.82\sqrt{0.1667 + \frac{9}{630}}$$

$$= 24.82\sqrt{0.1667 + 0.0143}$$

$$= 24.82\sqrt{0.1810}$$

$$= 24.82(0.4254)$$

$$= 10.56$$

We are now ready to put *confidence limits* around our predicted mean value. These have the same form which we have used repeatedly: estimator plus–minus so many estimated standard errors. In the present situation, the confidence interval is the following:

$$\hat{y}_k - t_* s_{\hat{y}_k} < \text{true mean } Y \text{ at } x = x_k < \hat{y}_k + t_* s_{\hat{y}_k}, \tag{8.6.6}$$

where

$$\hat{y}_k = \bar{y} + b(x - x_k),$$

$$s_{\hat{y}_k} = s\sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x - \bar{x})^2}},$$

$$s = \sqrt{\text{residual mean square in ANOVA}},$$

$$t_* = \text{the confidence-limit value of } t$$
$$\text{with } n - 2 \text{ d.f.}$$

In our example we have, from earlier calculations:

$$x_k = 24; \qquad \hat{y}_k = 174.30; \qquad s_{\hat{y}_k} = 10.56; \qquad n = 6$$

For a 95 percent confidence interval, we take t_* from the table of t with $6 - 2 = 4$ d.f., getting $t_* = 2.776$, and so have the confidence limits

$$174.30 \pm 2.776(10.56) = 174.30 \pm 29.31,$$

giving the 95 percent *confidence interval* as

$$144.9 < \text{true mean cost at 24 months} < 203.61,$$

or, more reasonable to report,

$$\$145 < \text{true mean cost at 24 months} < \$204.$$

Notice in formula (8.6.3) for the variance of the predicted mean value $\hat{y}_k$ that the variance of the slope **b** is always multiplied by $(x_k - \bar{x})^2$. That multiplier is zero if $x_k = \bar{x}$, and gets bigger and bigger as x_k is taken farther and farther away from $\bar{x}$. In other words, the further you go from $(\bar{x}, \bar{y})$ to predict the response, the more variation or wobble due to the slope will affect the prediction. Consider Figure 8.6.1.
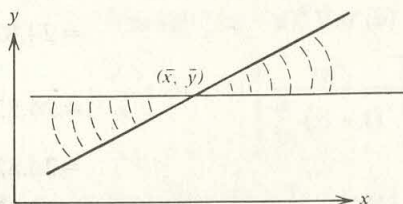
FIGURE 8.6.1

If the variation in the slope is depicted by two lines intersecting at $\bar{y}$, then the further you go from $\bar{x}$ in either direction, the larger the effect of the variation in slope.

Let's now calculate the standard error of the predicted value of the mean maintenance cost for each of the values of x_i in the data of our example. The following table gives the results. We have the values of $\hat{y}_i$ from Table 8.3.3 and set up the computation of Var $(\hat{y}_i)$ from (8.6.4) put into the convenient form:

$$\frac{s^2}{n} + \frac{s^2(x_k - 21)^2}{\sum (x_i - \bar{x})^2} = \frac{616.075}{6} + \frac{616.075}{630}(x_k - 21)^2$$

$$= 102.68 + 0.9779(x_k - 21)^2$$

From Table 8.6.1 we can quickly put 95 percent confidence limits around each $\hat{y}_i$, applying (8.6.6) and taking $t_* = 2.776$ as we did before in our example. (See Table 8.6.2.)

TABLE 8.6.1

x_i	$x_i - \bar{x}$	y_i	$\hat{y}_i$	$\hat{V}ar\,(\hat{y}_i)$	$s_{\hat{y}_i}$
6	−15	50	28.5	$102.68 + 0.9779(-15)^2 = 322.71$	17.96
12	−9	75	77.1	$102.68 + 0.9779(-9)^2 = 181.89$	13.49
18	−3	100	125.7	$102.68 + 0.9779(-3)^2 = 111.48$	10.56
24	+3	175	174.3	$102.68 + 0.9779(+3)^2 = 111.48$	10.56
30	+9	200	222.9	$102.68 + 0.9779(+9)^2 = 181.89$	13.49
36	+15	300	271.5	$102.68 + 0.9779(+15)^2 = 322.71$	17.96

TABLE 8.6.2

x_i	y_i	$\hat{y}_i$	$s_{\hat{y}_i}$	$t_* \cdot s_{\hat{y}_i}$	95 percent Confidence Limits	
					Lower	Upper
6	50	28.5	17.96	49.9	0*	78.4
12	75	77.1	13.49	37.4	39.7	114.5
18	100	125.7	10.56	29.3	96.4	155.0
24	175	174.3	10.56	29.3	145.0	203.6
30	200	222.9	13.49	37.4	185.5	260.3
36	300	271.5	17.96	49.9	221.6	321.4

*Lower bound must be zero for these data.

Plotting these confidence limits on the graph with the best-fitting straight line, we have Figure 8.6.2.

Notice how the confidence band widens as we move farther and farther from $x = \bar{x} = 21$. Outside the range of x that we have had in our actual data (here 6–36) even the very wide confidence band is undependable, and we must follow the scientist's general rule: extrapolation beyond the range of your data is extremely dangerous. Take care!
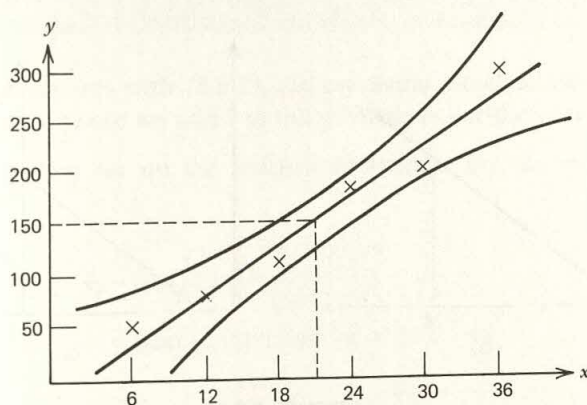
FIGURE 8.6.2

8.7 PREDICTING THE NEXT OBSERVATION AT A GIVEN VALUE OF x, SAY x_k

In the last section we dealt with the prediction of the mean response at a particular value of x. In our example we raised the question, "for cars 24 months of age, what is our prediction of mean maintenance cost, and what is a 95 percent confidence limit on that prediction?" Our answer to this used the regression analysis worked out thus far. Using the observed data and assuming that maintenance costs were linearly related to the age of cars, our prediction for the mean 6-month maintenance costs for cars 24 months old was $174.30. Our 95 percent confidence statement was that the true mean 6-month maintenance bill for cars 24 months old is between $145 and $204.

The natural question to ask is, "while that's not so bad for average maintenance costs, what about my particular car?" In other words, what can we say about an *individual* car or observation? Obviously when we start concerning ourselves about a single observation (here, a car), our point estimate will be the same but the car-to-car maintenance cost variability for cars of the same age is much bigger. Let's look at the following graph (Figure 8.7.1) predicting $\hat{y}_k$ when $x = x_k = 24$ months.
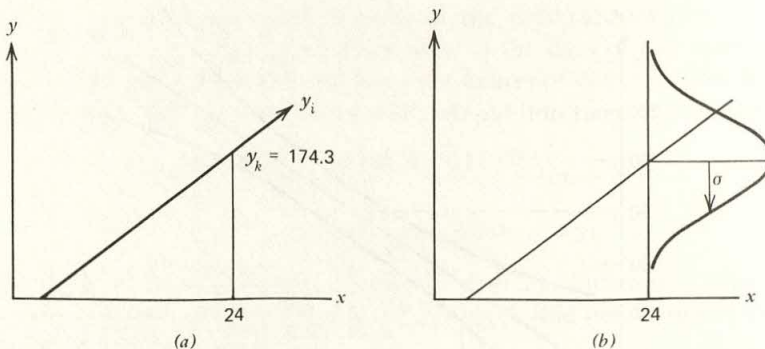
FIGURE 8.7.1

In Figure 8.7.1a we have the predicted value, $y_k = \$174.30$ when $x_k = 24$ months. From the preceding section we know that the estimate y_k has a variance based on the straight-line model; that is, its variance has two components, one due to estimating the overall mean by $\bar{y}$, and the other due to estimating the slope β by use of b. Hence our prediction of the regression line is subject to variability, and in Figure 8.6.2 we saw the shape of a confidence band for it. All this has to do with just the *mean* response at a given value of x. In Figure 8.7.1b we see an additional variance component. This component is due to the distribution of *single* observations about the regression line when $x_k = 24$.

Thus if we are to make a confidence interval statement about the next single observation, say at $x_k = 24$, we will need to add a component for the individual variance to the components due to mean and slope. Thus the variance of the next observation taken at x_k is found by adding σ^2 to the variance in (8.6.3):

$$\text{Var (next observation at } x_k) = \sigma^2 + \frac{\sigma^2}{n} + (x_k - \bar{x})^2 \left[\frac{\sigma^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \right] \qquad (8.7.1)$$

Substituting s^2 for σ^2 and taking the square root, we get the estimate of the standard deviation (standard error) of the next observation at $x = x_k$:

$$\left. \begin{array}{l} \text{Estimate of the} \\ \text{standard deviation of} \\ \text{the next observation} \\ \text{at } x = x_k \end{array} \right\} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}} \qquad (8.7.2)$$

Note how this compares with (8.6.5), the estimated standard deviation of the mean response at x_k: here we add 1 to the quantity under the square-root sign.

From this we then set up the confidence interval for the next individual observation on Y:

$$\hat{y}_k - t_{*}s \sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

$$< \text{next observation on } Y \text{ at } x = x_k$$

$$< \hat{y}_k + t_{*}s \sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum (x - \bar{x})^2}}, \qquad (8.7.3)$$

where the notation has the same meaning as in (8.6.6),

In our example on car maintenance costs for car age 24 months, we had earlier:

$$\hat{y}_k = 174.30, \qquad s_{\hat{y}_k} = 24.82\sqrt{0.1810} = 24.82(0.4254) = 10.56$$

To predict the 6-month maintenance cost of a *specific individual* car 24 months old, we again take the point estimate as $\hat{y}_k = 174.30$. But now we use (8.7.2) and (8.7.3). The standard error is:

$$24.82\sqrt{1+0.1810} = 24.82\sqrt{1.1810} = 24.82(1.087) = 26.98,$$

and then the limits of a 95 percent confidence interval are

$$174.30 \pm (2.776)(26.98) = 174.30 \pm 74.90,$$

giving the 95 percent confidence interval:

$$99.40 < \left\{ \begin{array}{c} \text{next observation on} \\ \text{6-month maintenance cost} \\ \text{of car 24 months old} \end{array} \right\} < 249.20$$

The 95 percent confidence limits on the next observation at each of the x values given by the x_i of our data are recorded in the Table 8.7.1, where also the limits in Table 8.6.2 have been repeated for comparison.

TABLE 8.7.1

x_i	y_i	$\hat{y}_i$	Standard Deviation of Mean Response	Standard Deviation of Next Observation	95 percent Confidence Limits			
					True Mean Response		Next Observation	
					low	high	low	high
6	50	28.5	17.96	30.68	0[a]	78.4	0[a]	113.7
12	75	77.1	13.49	28.25	39.7	114.5	0[a]	155.5
18	100	125.7	10.56	26.98	96.4	155.0	50.8	200.6
24	175	174.3	10.56	26.98	145.0	203.6	99.4	249.2 [b]
30	200	222.9	13.49	28.25	185.5	260.3	144.5	301.3
36	300	271.5	17.96	30.68	221.6	321.4	186.3	356.7

[a] As in Table 8.6.2, zero cost must be taken as lowest possible bound.

[b] Calculations for this interval are shown in the text above.

As we expected, the confidence intervals on the *next observation* are much wider than those on the *mean* response, reflecting the additional variance component due to the distribution of individual responses about the true regression line.

8.8 ANALYSIS OF RESIDUALS

Let's consider what we've done up to this point in this chapter. We started out by using a simple predictive model $\hat{y}_i = \bar{y}$ or $\hat{y}_i = 150$. Then we noted that the deviations $\hat{y}_i - 150$ were closely related to the values of an independent or predictive variable x_i. So we then expanded our model through least squares and obtained the predictive model $\hat{y}_i = 150 + 8.10(x_i - 21)$. We found that this model explained 94.4 percent of the remaining residual variation. This straight-line model is a good one. To reinforce this, using a t test we rejected the null hypothesis that β, the true slope, was equal to zero. Then in order to use the model for prediction, we calculated confidence intervals for both the mean response at a given value of x and for the next observation at a given value of x.

Even though everything we have done has led to evidence of an excellent linear prediction model, we should still make sure that there exists no evidence that this *good* model could be further improved. As we did in Step 2, Section 8.3, let us look at the residuals in Table 8.3.3, and also divide them by the residual standard deviation $\sqrt{s^2}$ from the ANOVA Table 8.3.4, and retabulate them (Table 8.8.1).

TABLE 8.8.1

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)/s$
6	50	28.5	21.5	21.5/24.82 = .87
12	75	77.1	-2.1	-2.1/24.82 = -.08
18	100	125.7	-25.7	-25.7/24.82 = -1.04
24	175	174.3	0.7	0.7/24.82 = .03
30	200	222.9	-22.9	-22.9/24.82 = -.92
36	300	271.5	28.5	28.5/24.82 = 1.15

There are no discernible patterns in the residuals, and, further, none of the standardized residuals is close to ± 1.96. With this additional evidence we should now feel much more secure in using the predictive model,

$$\hat{y} = 150 + 8.10(x - 21).$$

The final lesson then is: no matter how statistically significant any model seems to be, *always* look at the residuals after fitting the model. Then, and only then, should one be secure with a predictive model.

EXERCISES

8.8.1 Using the table of calculations shown in Exercise 8.3.8 and the best-fitting straight line $\hat{y} = 10 - 0.82(x - 7)$:

    a. Determine the standard deviation of the predicted value $\hat{y}_k$ when $x_k = 7$ (Hint: use format in Table 8.6.1).

    b. Determine the standard deviation of the predicted value of $\hat{y}_k$ when $x_k = 13$.

    c. Why is the value of $s_{\hat{y}_k}$ larger when $x_k = 13$ than when $x_k = 7$?

8.8.2 Using the same table of calculations shown in Exercise 8.3.8,

    a. Calculate $s_{\hat{y}_k}$ when $x_k = 1$.

    b. Using format of Table 8.6.2, construct 95 percent confidence limits on the average response of y, using a linear model, at the three points where

$$x_k = 1, \qquad x_k = 7, \qquad \text{and} \qquad x_k = 13.$$

    c. What value of t_* did you use? Why?

    d. Plot the 95 percent confidence limits on a graph relating y and x.

8.8.3 For the same calculations in Exercise 8.8.2, add the next columns such that one obtains the 95 percent confidence limits on the next observation.

8.8.4 Using the data in Exercise 8.3.2 (also used in Exercises 8.3.4 and 8.3.6) and using the format of Tables 8.6.1, 8.6.2, and 8.7.1, construct the 95 percent confidence limits on the next observation for each value of x_k in the data set. (Hint: in order to do this, you'll need many of the results previously calculated.)

8.8.5 Calculate the standardized residuals for the fit of the data in Exercise 8.3.4, using the format shown in Table 8.8.1.

    a. Are there any discernible patterns in the data?

    b. What are your final conclusions about the fitting of the linear model to these data?

8.8.6 In Exercise 8.3.7, take the residuals obtained in this exercise and standardize them as indicated in Table 8.8.1.

    a. Does this added information reinforce your opinion of the usefulness of the linear model?

    b. What is your opinion about the utility of this linear prediction equation?

8.9 SUMMARY EXERCISES

8.9.1 During the time period 1958–1961, the shipments of the machinery and equipment industries to customers began to rise. The quarterly shipments during this period are shown below:

	Quarter	Shipments (billions of 1957–1959 dollars)
	x	y
January–March, 1958:	1	7.5
	2	7.4
	3	7.5
	4	7.7
	5	8.0
	6	8.3
	7	8.5
	8	8.4
	9	8.4
	10	8.6
	11	8.4
	12	8.1
	13	8.0
	14	8.2
	15	8.5

a. Fit a linear model of the form $y_i = \beta_0 + \beta x_i + E_i$ to these data.

b. Test the hypothesis $H_0: \beta = 0$ versus $H_A: \beta \neq 0$, using $\alpha = 5$ percent. Draw tentative conclusions.

c. How good a fit is this model?

d. After finding the best-fitting straight line, calculate the residuals from the fit.

e. Are you satisfied with this model? Do you think the model could be improved? Why, or why not?

8.9.2 The U.S. Coast Guard is responsible for dealing with the oil spilled in the harbors of the U.S. Under normal conditions (no major accidents), oil in the harbor is caused by leakages, and discharges from ships, plants, and so on. Much of the oil that can be detected by eye comes from unknown sources, and much of the oil spillage is often not detected. It has been conjectured that the amount of oil spillage reported is a direct function of how many ships are boarded by the Coast Guard and inspected for oil-spillage potential (i.e., the

more you look for oil trouble, the more you find). The following data were collected on this conjecture:

	Average Number of Ship Boardings per District by Quarter x	Average District Quarterly Oil Discharge Volume y
January–March, 1971	723	86,888
April–June	625	26,844
July–September	620	45,975
October–December	612	49,211
January–March, 1972	712	88,876
April–June	565	44,652
July–September	517	25,610
October–December	554	67,192

Using an $\alpha = 5$ percent significance level, do you agree or disagree with the conjecture?

8.9.3   A test was conducted to determine the capacity of a soap-making machine for different cooling-water temperatures. The results were as follows for a fixed water flow rate:

Cooling-Water Temperature in degrees Fahrenheit	Soap-Production Rate in pounds per minute
63	185
63	170
64	160
64	150
64	155
66	150
67	140
69	140
69	120
70	115
70	125
72	110

a. Draw your graphical estimate of the best regression line.
b. Calculate the regression line for the model $Y_i = \beta_0 + \beta x_i + E_i$, using the least-squares procedure.
c. Draw the calculated line for comparison with (a).
d. Is the regression significant at $\alpha = .05$?
e. Calculate 95 percent confidence limits on the expected (mean) response at $x_k = \bar{x}$.
f. Draw conclusions.

8.9.4   The problem of the availability of physicians is a serious one. People who live in remote or relatively unpopulated areas have difficulty reaching a physician. Several programs have been initiated by the government to alleviate the problem. One outstanding educator has suggested that establishing a medical school in a university well removed from the center of population would attract physicians to that area. The question arose, "excluding the county that includes a medical school, is there a relationship between the distance to a medical school from the county seat and the number of physicians in the county?" The following data are a sample of a complete set from the state of North Carolina.

County	Distance (Rounded to Nearest Mile) to a Medical School x	Number of Physicians per 100,000 Population y
1	84	19
2	38	58
3	77	58
4	32	79
5	45	72
6	100	98
7	56	110
8	45	117
9	28	95
10	25	130
11	147	138
12	95	84
13	82	55
14	63	81
15	18	169
16	65	165
17	10	40
18	79	72
19	52	16
20	67	227
21	77	123
22	36	215
23	76	100
24	47	110
25	48	117

a. Fit a straight line to the data.
b. What is your estimate of the slope β?
c. What is the estimated standard deviation of the slope?
d. Test $H_0: \beta = 0$ against $H_A: \beta \neq 0$, using $\alpha = 10$ percent.
e. What are your conclusions?

8.9.5 The following data have been collected on required homework grades and on the test performance of students in one science class in a major university. The instructor of the class decided to determine whether he could predict test performance using average required homework grades obtained prior to the test.

Required Homework Average Grade	Test-Performance Grade
53	30
67	33
51	56
58	48
64	40
65	39
69	46
62	53
70	58
78	55
77	65
84	64
93	58
88	68
74	83
93	73
92	80

a. Determine the best-fitting linear prediction equation of the form $\hat{y} = b_0 + bx$.
b. What is the estimated standard deviation of b?
c. Place 95 percent confidence limits on true β.
d. What is the estimated standard deviation of the predicted mean test score when the average homework score is $x_k = 70$?
e. By choosing three values of x, draw 95 percent confidence limits for the true mean test score line.
f. Draw conclusions.

8.9.6 The following data show the amount of money (millions of dollars) invested in public and private nonresidential construction. These data are in constant dollars, which eliminates the effect of price changes.

a. Determine the slope of the best-fitting straight line relating year to public nonresidential construction dollars. [Take year as $x = 1, 2, \ldots, 13$.]
b. Determine the slope of the best-fitting straight line relating private nonresidential construction (in dollars) to the year.
c. What can you say about these data?
d. Draw inferences with the results you have calculated.

	Public	Private
1960	15	16
1961	16	17
1962	16.5	17.5
1963	18	18
1964	19	19
1965	21	25
1966	22	27
1967	24	28
1968	26	30
1969	26.5	33.5
1970	27	35
1971	28	36
1972	29	39

8.9.7 "Sticky" shampoo shipments are given for months 4–23 following national introduction.

x = Months since Introduction	y = Shipment Size (thousands of cases)
4	43
5	55
6	73
7	53
8	101
9	59
10	43
11	52
12	44
13	44
14	51
15	53
16	73
17	81
18	68
19	50
20	76
21	97
22	82
23	53

$\sum x = 270$
$\sum x^2 = 4{,}310$
$n = 20$

$\sum y = 1{,}251$
$\sum y^2 = 84{,}261$
$\sum xy = 17{,}505$

a. Fit a straight line to these data.
b. What is the expected average monthly increase for shipments of "Sticky" shampoo? Is this significant at 10 percent risk?
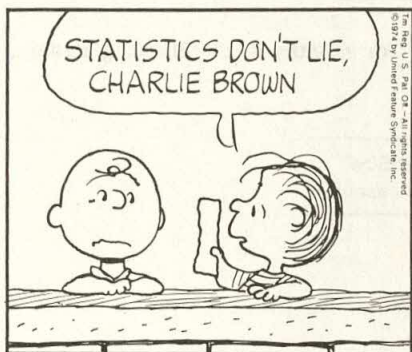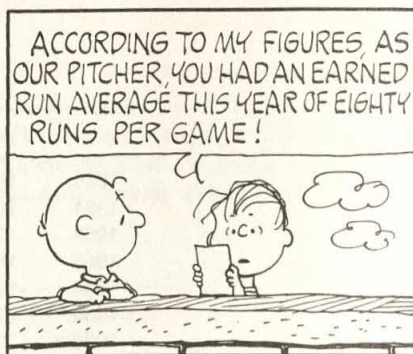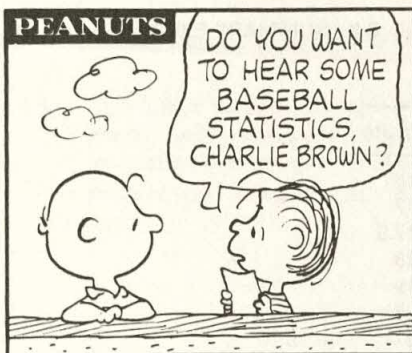
9

A Backward Glance

9

Mecca Community College students turned out to be an interesting bunch! Now that the material in this book has been read, digested, and, hopefully, learned, it might be interesting to you to find out something about the students in your own school.

As an exercise, take the questions asked of the Mecca students and conduct a study in your school. Conjecture on the results. Discuss the difficulties encountered in gathering the data. Comment on the problems of getting a random sample. After the data are collected, use the information you've learned about statistics and analyze the data. How does your school compare with Mecca Community College?

We, the authors, would welcome your data for future reference and for improving the examples in this book. When you put your statistics together, you may agree with Charlie Brown on the opposite page that "they sure shoot off their mouth a lot!" Good luck.

Appendix A
TABLES

TABLE A-1. Squares and Square Roots

N	N²	√N	√10N	N	N²	√N	√10N	N	N²	√N	√10N
100	10 000	10.0000	31.6228	150	22 500	12.2474	38.7298	200	40 000	14.1421	44.7214
101	10 210	10.0499	31.7805	151	22 801	12.2882	38.8587	201	40 401	14.1774	44.8330
102	10 404	10.0995	31.9374	152	23 104	12.3288	38.9872	202	40 804	14.2127	44.9444
103	10 609	10.1489	32.0936	153	23 409	12.3693	39.1152	203	41 209	14.2478	45.0555
104	10 816	10.1980	32.2490	154	23 716	12.4097	39.2428	204	41 616	14.2829	45.1664
105	11 025	10.2470	32.4037	155	24 025	12.4499	39.3700	205	42 025	14.3178	45.2769
106	11 236	10.2956	32.5576	156	24 336	12.4900	39.4968	206	42 436	14.3527	45.3872
107	11 449	10.3441	32.7109	157	24 649	12.5300	39.6232	207	42 849	14.3875	45.4973
108	11 664	10.3923	32.8634	158	24 964	12.5698	39.7492	208	43 264	14.4222	45.6070
109	11 881	10.4403	33.0151	159	25 281	12.6095	39.8748	209	43 681	14.4568	45.7165
110	12 100	10.4881	33.1662	160	25 600	12.6491	40.0000	210	44 100	14.4914	45.8258
111	12 321	10.5357	33.3167	161	25 921	12.6886	40.1248	211	44 521	14.5258	45.9347
112	12 544	10.5830	33.4664	162	26 244	12.7279	40.2492	212	44 944	14.5602	46.0435
113	12 769	10.6301	33.6155	163	26 569	12.7671	40.3733	213	45 369	14.5945	46.1519
114	12 996	10.6771	33.7639	164	26 896	12.8062	40.4969	214	45 796	14.6287	46.2601
115	13 225	10.7238	33.9116	165	27 225	12.8452	40.6202	215	46 225	14.6629	46.3681
116	13 456	10.7703	34.0588	166	27 556	12.8841	40.7431	216	46 656	14.6969	46.4758
117	13 689	10.8167	34.2053	167	27 889	12.9228	40.8656	217	47 089	14.7309	46.5833
118	13 924	10.8628	34.3511	168	28 224	12.9615	40.9878	218	47 524	14.7648	46.6905
119	14 161	10.9087	34.4964	169	28 561	13.0000	41.1096	219	47 961	14.7986	46.7974
120	14 400	10.9545	34.6410	170	29 900	13.0384	41.2311	220	48 400	14.8324	46.9042
121	14 641	11.0000	34.7851	171	29 241	13.0767	41.3521	221	48 841	14.8661	47.0106
122	14 884	11.0454	34.9285	172	29 584	13.1149	41.4729	222	49 284	14.8997	47.1169
123	15 129	11.0905	35.0714	173	29 929	13.1529	41.5933	223	49 729	14.9332	47.2229
124	15 376	11.1355	35.2136	174	30 276	13.1909	41.7133	224	50 176	14.9666	47.3286
125	15 625	11.1803	35.3553	175	30 625	13.2288	41.8330	225	50 625	15.0000	47.4342
126	15 876	11.2250	35.4965	176	30 976	13.2665	41.9524	226	51 076	15.0333	47.5395
127	16 129	11.2694	35.6371	177	31 329	13.3041	42.0714	227	51 529	15.0665	47.6445
128	16 384	11.3137	35.7771	178	31 684	13.3417	42.1900	228	51 984	15.0997	47.7493
129	16 641	11.3578	35.9166	179	32 041	13.3791	42.3084	229	52 441	15.1327	47.8539
130	16 900	11.4018	36.0555	180	32 400	13.4164	42.4264	230	52 900	15.1658	47.9583
131	17 161	11.4455	36.1939	181	32 761	13.4536	42.5441	231	53 361	15.1987	48.0625
132	17 424	11.4891	36.3318	182	33 124	13.4907	42.6615	232	53 824	15.2315	48.1664
133	17 689	11.5326	36.4692	183	33 489	13.5277	42.7785	233	54 289	15.2643	48.2701
134	17 956	11.5758	36.6060	184	33 856	13.5647	42.8952	234	54 756	15.2971	48.3735
135	18 225	11.6190	36.7423	185	34 225	13.6015	43.0116	235	55 225	15.3297	48.4768
136	18 496	11.6619	36.8782	186	34 596	13.6382	43.1277	236	55 696	15.3623	48.5798
137	18 769	11.7047	37.0135	187	34 969	13.6748	43.2435	237	56 169	15.3948	48.6826
138	19 044	11.7473	37.1484	188	35 344	13.7113	43.3590	238	56 644	15.4272	48.7852
139	19 321	11.7898	37.2827	189	35 721	13.7477	43.4741	239	57 121	15.4596	48.8876
140	19 600	11.8322	37.4166	190	36 100	13.7840	43.5890	240	57 600	15 4919	48.9898
141	19 881	11.8743	37.5500	191	36 481	13.8203	43.7035	241	58 081	15.5242	49.0918
142	20 164	11.9164	37.6829	192	36 864	13.8564	43.8178	242	58 564	15.5563	49.1935
143	20 449	11.9583	37.8153	193	37 249	13.8924	43.9318	243	59 049	15.5885	49.2950
144	20 736	12.0000	37.9473	194	37 636	13.9284	44.0454	244	59 536	15.6205	49.3964
145	21 025	12.0416	38.0789	195	38 025	13.9642	44.1588	245	60 025	15.6525	49.4975
146	21 316	12.0830	38.2099	196	38 416	14.0000	44.2719	246	60 516	15.6844	49.5984
147	21 609	12.1244	38.3406	197	38 809	14.0357	44.3847	247	61 009	15.7162	49.6991
148	21 904	12.1655	38.4708	198	39 204	14.0712	44.4972	248	61 504	15.7480	49.7996
149	22 201	12.2066	38.6005	199	39 601	14.1067	44.6094	249	62 001	15.7797	49.8999
150	22 500	12.2474	38.7298	200	40 000	14.1421	44.7214	250	62 500	15.8114	50.0000

N	N²	√N	√10N	N	N²	√N	√10N	N	N²	√N	√10N
250	62 500	15.8114	50.0000	300	90 000	17.3205	54.7723	350	122 500	18.7083	59.1608
251	63 001	15.8430	50.0999	301	90 601	17.3493	54.8635	351	123 201	18.7350	59.2453
252	63 504	15.8745	50.1996	302	91 204	17.3781	54.9545	352	123 904	18.7617	59.3296
253	64 009	15.9060	50.2991	303	91 809	17.4069	55.0454	353	124 609	18.7883	59.4138
254	64 516	15.9374	50.3984	304	92 416	17.4356	55.1362	354	125 316	18.8149	59.4979
255	65 025	15.9687	50.4975	305	93 025	17.4642	55.2268	355	126 025	18.8414	59.5819
256	65 536	16.0000	50.5964	306	93 636	17.4929	55.3173	356	126 736	18.8680	59.6657
257	66 049	16.0312	50.6952	307	94 249	17.5214	55.4076	357	127 449	18.8944	59.7495
258	66 564	16.0624	50.7937	308	94 864	17.5499	55.4977	358	128 164	18.9209	59.8331
259	67 081	16.0935	50.8920	309	95 481	17.5784	55.5878	359	128 881	18.9473	59.9166
260	67 600	16.1245	50.9902	310	96 100	17.6068	55.6776	360	129 600	18.9737	60.0000
261	68 121	16.1555	51.0882	311	96 721	17.6352	55.7674	361	130 321	19.0000	60.0833
262	68 644	16.1864	51.1859	312	97 344	17.6635	55.8570	362	131 044	19.0263	60.1664
263	69 169	16.2173	51.2835	313	97 969	17.6918	55.9464	363	131 769	19.0526	60.2495
264	69 696	16.2481	51.3809	314	98 596	17.7200	56.0357	364	132 496	19.0788	60.3324
265	70 225	16.2788	51.4782	315	99 225	17.7482	56.1249	365	133 225	19.1050	60.4152
266	70 756	16.3095	51.5752	316	99 856	17.7764	56.2139	366	133 956	19.1311	60.4979
267	71 289	16.3401	51.6720	317	100 489	17.8045	56.3028	367	134 689	19.1572	60.5805
268	71 824	16.3707	51.7687	318	101 124	17.8326	56.3915	368	135 424	19.1833	60.6630
269	72 361	16.4012	51.8652	319	101 761	17.8606	56.4801	369	136 161	19.2094	60.7454
270	72 900	16.4317	51.9615	320	102 400	17.8885	56.5685	370	136 900	19.2354	60.8276
271	73 441	16.4621	52.0577	321	103 041	17.9165	56.6569	371	137 641	19.2614	60.9098
272	73 984	16.4924	52.1536	322	103 684	17.9444	56.7450	372	138 384	19.2873	60.9918
273	74 529	16.5227	52.2494	323	104 329	17.9722	56.8331	373	139 129	19.3132	61.0737
274	75 076	16.5529	52.3450	324	104 976	18.0000	56.9210	374	139 876	19.3391	61.1555
275	75 625	16.5831	52.4404	325	105 625	18.0278	57.0088	375	140 625	19.3649	61.2372
276	76 176	16.6132	52.5357	326	106 276	18.0555	57.0964	376	141 376	19.3907	61.3188
277	76 729	16.6433	52.6308	327	106 929	18.0831	57.1839	377	142 129	19.4165	61.4003
278	77 284	16.6733	52.7257	328	107 584	18.1108	57.2713	378	142 884	19.4422	61.4817
279	77 841	16.7033	52.8205	329	108 241	18.1384	57.3585	379	143 641	19.4679	61.5630
280	78 400	16.7332	52.9150	330	108 900	18.1659	57.4456	380	144 400	19.4936	61.6441
281	78 961	16.7631	53.0094	331	109 561	18.1934	57.5326	381	145 161	19.5192	61.7252
282	79 524	16.7929	53.1037	332	110 224	18.2209	57.6194	382	145 924	19.5448	61.8061
283	80 089	16.8226	53.1977	333	110 889	18.2483	57.7062	383	146 689	19.5704	61.8870
284	80 656	16.8523	53.2917	334	111 556	18.2757	57.7927	384	147 456	19.5959	61.9677
285	81 225	16.8819	53.3854	335	112 225	18.3030	57.8792	385	148 225	19.6214	62.0484
286	81 796	16.9115	53.4790	336	112 896	18.3303	57.9655	386	148 996	19.6469	62.1289
287	82 369	16.9411	53.5724	337	113 569	18.3576	58.0517	387	149 769	19.6723	62.2093
288	82 944	16.9706	53.6656	338	114 244	18.3848	58.1378	388	150 544	19.6977	62.2896
289	83 521	17.0000	53.7587	339	114 921	18.4120	58.2237	389	151 321	19.7231	62.3699
290	84 100	17.0294	53.8516	340	115 600	18.4391	58.3095	390	152 100	19.7484	62.4500
291	84 681	17.0587	53.9444	341	116 281	18.4662	58.3952	391	152 881	19.7737	62.5300
292	85 264	17.0880	54.0370	342	116 964	18.4932	58.4808	392	153 664	19.7990	62.6099
293	85 849	17.1172	54.1295	343	117 649	18.5203	58.5662	393	154 449	19.8242	62.6897
294	86 436	17.1464	54.2218	344	118 336	18.5472	58.6515	394	155 236	19.8494	62.7694
295	87 025	17.1756	54.3139	345	119 025	18.5742	58.7367	395	156 025	19.8746	62.8490
296	87 616	17.2047	54.4059	346	119 716	18.6011	58.8218	396	156 816	19.8997	62.9285
297	88 209	17.2337	54.4977	347	120 409	18.6279	58.9067	397	157 609	19.9249	63.0079
298	88 804	17.2627	54.5894	348	121 104	18.6548	58.9915	398	158 404	19.9499	63.0872
299	89 401	17.2916	54.6809	349	121 801	18.6815	59.0762	399	159 201	19.9750	63.1664
300	90 000	17.3205	54.7723	350	122 500	18.7083	59.1608	400	160 000	20.0000	63.2456

N	N²	√N	√10N	N	N²	√N	√10N	N	N²	√N	√10N
400	160 000	20.0000	63.2456	450	202 500,	21.2132	67.0820	500	250 000	22.3607	70.7107
401	160 801	20.0250	63.3246	451	203 401	21.2368	67.1565	501	251 001	22.3830	70.7814
402	161 604	20.0499	63.4035	452	204 304	21.2603	67.2309	502	252 004	22.4054	70.8520
403	162 409	20.0749	63.4823	453	205 209	21.2838	67.3053	503	253 009	22.4277	70.9225
404	163 216	20.0998	63.5610	454	206 116	21.3073	67.3795	504	254 016	22.4499	70.9930
405	164 025	20.1246	63.6396	455	207 025	21.3307	67.4537	505	255 025	22.4722	71.0634
406	164 836	20.1494	63.7181	456	207 936	21.3542	67.5278	506	256 036	22.4944	71.1337
407	165 649	20.1742	63.7966	457	208 849	21.3776	67.6018	507	257 049	22.5167	71.2039
408	166 464	20.1990	63.8749	458	209 764	21.4009	67.6757	508	258 064	22.5389	71.2741
409	167 281	20.2237	63.9531	459	210 681	21.4243	67.7495	509	259 081	22.5610	71.3442
410	168 100	20.2485	64.0312	460	211 600	21.4476	67.8233	510	260 100	22.5832	71.4143
411	168 921	20.2731	64.1093	461	212 521	21.4709	67.8970	511	261 121	22.6053	71.4843
412	169 744	20.2978	64.1872	462	213 444	21.4942	67.9706	512	262 144	22.6274	71.5542
413	170 569	20.3224	64.2651	463	214 369	21.5174	68.0441	513	263 169	22.6495	71.6240
414	171 396	20.3470	64.3428	464	215 296	21.5407	68.1175	514	264 196	22.6716	71.6938
415	172 225	20.3715	64.4205	465	216 225	21.5639	68.1909	515	265 225	22.6936	71.7635
416	173 056	20.3961	64.4981	466	217 156	21.5870	68.2642	516	266 256	22.7156	71.8331
417	173 889	20.4206	64.5755	467	218 089	21.6102	68.3374	517	267 289	22.7376	71.9027
418	174 724	20.4450	64.6529	468	219 024	21.6333	68.4105	518	268 324	22.7596	71.9722
419	175 561	20.4695	64.7302	469	219 961	21.6564	68.4836	519	269 361	22.7816	72.0417
420	176 400	20.4939	64.8074	470	220 900	21.6795	68.5565	520	270 400	22.8035	72.1110
421	177 241	20.5183	64.8845	471	221 841	21.7025	68.6294	521	271 441	22.8254	72.1803
422	178 084	20.5426	64.9615	472	222 784	21.7256	68.7023	522	272 484	22.8473	72.2496
423	178 929	20.5670	65.0385	473	223 729	21.7486	68.7750	523	273 529	22.8692	72.3187
424	179 776	20.5913	65.1153	474	224 676	21.7715	68.8477	524	274 576	22.8910	72.3878
425	180 625	20.6155	65.1920	475	225 625	21.7945	68.9202	525	275 625	22.9129	72.4569
426	181 476	20.6398	65.2687	476	226 576	21.8174	68.9928	526	276 676	22.9347	72.5259
427	182 329	20.6640	65.3452	477	227 529	21.8403	69.0652	527	277 729	22.9565	72.5948
428	183 184	20.6882	65.4217	478	228 484	21.8632	69.1375	528	278 784	22.9783	72.6636
429	184 041	20.7123	65.4981	479	229 441	21.8861	69.2098	529	279 841	23.0000	72.7324
430	184 900	20.7364	65.5744	480	230 400	21.9089	69.2820	530	280 900	23.0217	72.8011
431	185 761	20.7605	65.6506	481	231 361	21.9317	69.3542	531	281 961	23.0434	72.8697
432	186 624	20.7846	65.7267	482	232 324	21.9545	69.4252	532	283 024	23.0651	72.9383
433	187 489	20.8087	65.8027	483	233 289	21.9773	69.4982	533	284 089	23.0868	73.0068
434	188 356	20.8327	65.8787	484	234 256	22.0000	69.5701	534	285 156	23.1084	73.0753
435	189 225	20.8567	65.9545	485	235 225	22.0227	69.6419	535	286 225	23.1301	73.1437
436	190 096	20.8806	66.0303	486	236 196	22.0454	69.7137	536	287 296	23.1517	73.2120
437	190 969	20.9045	66.1060	487	237 169	22.0681	69.7854	537	288 369	23.1733	73.2803
438	191 844	20.9284	66.1816	488	238 144	22.0907	69.8570	538	289 444	23.1948	73.3485
439	192 721	20.9523	66.2571	489	239 121	22.1133	69.9285	539	290 521	23.2164	73.4166
440	193 600	20.9762	66.3325	490	240 100	22.1359	70.0000	540	291 600	23.2379	73.4847
441	194 481	21.0000	66.4078	491	241 081	22.1585	70.0714	541	292 681	23.2594	73.5527
442	195 364	21.0238	66.4831	492	242 064	22.1811	70.1427	542	293 764	23.2809	73.6206
443	196 249	21.0476	66.5582	493	243 049	22.2036	70.2140	543	294 849	23.3024	73.6885
444	197 136	21.0713	66.6333	494	244 036	22.2261	70.2851	544	295 936	23.3238	73.7564
445	198 025	21.0950	66.7083	495	245 025	22.2486	70.3562	545	297 025	23.3452	73.8241
446	198 916	21.1187	66.7832	496	246 016	22.2711	70.4273	546	298 116	23.3666	73.8918
447	199 809	21.1424	66.8581	497	247 009	22.2935	70.4982	547	299 209	23.3880	73.9594
448	200 704	21.1660	66.9328	498	248 004	22.3159	70.5691	548	300 304	23.4094	74.0270
449	201 601	21.1896	67.0075	499	249 001	22.3383	70.6399	549	301 401	23.4307	74.0945
450	202 500	21.2132	67.0820	500	250 000	22.3607	70.7107	550	302 500	23.4521	74.1620

N	N²	√N	√10N	N	N²	√N	√10N	N	N²	√N	√10N
550	302 500	23.4521	74.1620	600	360 000	24.4949	77.4597	650	422 500	25.4951	80.6226
551	303 601	23.4734	74.2294	601	361 201	24.5153	77.5242	651	423 801	25.5147	80.6846
552	304 704	23.4947	74.2967	602	362 404	24.5357	77.5887	652	425 104	25.5343	80.7465
553	305 809	23.5160	74.3640	603	363 609	24.5561	77.6531	653	426 409	25.5539	80.8084
554	306 916	23.5372	74.4312	604	364 816	24.5764	77.7174	654	427 716	25.5734	80.8703
555	308 025	23.5584	74.4983	605	366 025	24.5967	77.7817	655	429 025	25.5930	80.9321
556	309 136	23.5797	74.5654	606	367 236	24.6171	77.8460	656	430 336	25.6125	80.9938
557	310 249	23.6008	74.6324	607	368 449	24.6374	77.9102	657	431 649	25.6320	81.0555
558	311 364	23.6220	74.6994	608	369 664	24.6577	77.9744	658	432 964	25.6515	81.1172
559	312 481	23.6432	74.7663	609	370 881	24.6779	78.0385	659	434 281	25.6710	81.1788
560	313 600	23.6643	74.8331	610	372 100	24.6982	78.1025	660	435 600	25.6905	81.2404
561	314 721	23.6854	74.8999	611	373 321	24.7184	78.1665	661	436 921	25.7099	81.3019
562	315 844	23.7065	74.9667	612	374 544	24.7386	78.2304	662	438 244	25.7294	81.3634
563	316 969	23.7276	75.0333	613	375 769	24.7588	78.2943	663	439 569	25.7488	81.4248
564	318 096	23.7487	75.0999	614	376 996	24.7790	78.3572	664	440 896	25.7682	81.4862
565	319 225	23.7697	75.1665	615	378 225	24.7992	78.4219	665	442 225	25.7876	81.5475
566	320 356	23.7908	75.2330	616	379 456	24.8193	78.4857	666	443 556	25.8070	81.6088
567	321 489	23.8118	75.2994	617	380 689	24.8395	78.5493	667	444 889	25.8263	81.6701
568	322 624	23.8328	75.3658	618	381 924	24.8596	78.6130	668	446 224	25.8457	81.7313
569	323 761	23.8537	75.4321	619	383 161	24.8797	78.6766	669	447 561	25.8650	81.7924
570	324 900	23.8747	75.4983	620	384 400	24.8998	78.7401	670	448 900	25.8844	81.8535
571	326 041	23.8956	75.5645	621	385 641	24.9199	78.8036	671	450 241	25.9037	81.9146
572	327 184	23.9165	75.6307	622	386 884	24.9399	78.8670	672	451 584	25.9230	81.9756
573	328 329	23.9374	75.6968	623	388 129	24.9600	78.9303	673	452 929	25.9422	82.0366
574	329 476	23.9583	75.7628	624	389 376	24.9800	78.9937	674	454 276	25.9615	82.0975
575	330 625	23.9792	75.8288	625	390 625	25.0000	79.0569	675	455 625	25.9808	82.1584
576	331 776	24.0000	75.8947	626	391 876	25.0200	79.1202	676	456 976	26.0000	82.2192
577	332 929	24.0208	75.9605	627	393 129	25.0400	79.1833	677	458 329	26.0192	82.2800
578	334 084	24.0416	76.0263	628	394 384	25.0599	79.2465	678	459 684	26.0384	82.3408
579	335 241	24.0624	76.0920	629	395 641	25.0799	79.3095	679	461 041	26.0576	82.4015
580	336 400	24.0832	76.1577	630	396 900	25.0998	79.3725	680	462 400	26.0768	82.4621
581	337 561	24.1039	76.2234	631	398 161	25.1197	79.4355	681	463 761	26.0960	82.5227
582	338 724	24.1247	76.2889	632	399 424	25.1396	79.4984	682	465 124	26.1151	82.5833
583	339 889	24.1454	76.3544	633	400 689	25.1595	79.5613	683	466 489	26.1343	82.6438
584	341 056	24.1661	76.4199	634	401 956	25.1794	79.6241	684	467 856	26.1534	82.7043
585	342 225	24.1868	76.4853	635	403 225	25.1992	79.6869	685	469 225	26.1725	82.7647
586	343 396	24.2074	76.5506	636	404 496	25.2190	79.7496	686	470 596	26.1916	82.8251
587	344 569	24.2281	76.6159	637	405 769	25.2389	79.8123	687	471 969	26.2107	82.8855
588	345 744	24.2487	76.6812	638	407 044	25.2587	79.8749	688	473 344	26.2298	82.9458
589	346 921	24.2693	76.7463	639	408 321	25.2784	79.9375	689	474 721	26.2488	83.0060
590	348 100	24.2899	76.8115	640	409 600	25.2982	80.0000	690	476 100	26.2679	83.0662
591	349 281	24.3105	76.8765	641	410 881	25.3180	80.0625	691	477 481	26.2869	83.1264
592	350 464	24.3311	76.9415	642	412 164	25.3377	80.1249	692	478 864	26.3059	83.1865
593	351 649	24.3516	77.0065	643	413 449	25.3574	80.1873	693	480 249	26.3249	83.2466
594	352 836	24.3721	77.0714	644	414 736	25.3772	80.2496	694	481 636	26.3439	83.3067
595	354 025	24.3926	77.1362	645	416 025	25.3969	80.3119	695	483 025	26.3629	83.3667
596	355 216	24.4131	77.2010	646	417 316	25.4165	80.3741	696	484 416	26.3818	83.4266
597	356 409	24.4336	77.2658	647	418 609	25.4362	80.4362	697	485 809	26 4008	83.4865
598	357 604	24.4540	77.3305	648	419 904	25.4558	80.4984	698	487 204	26.4197	83.5464
599	358 801	24.4745	77.3951	649	421 201	25.4755	80.5605	699	488 601	26.4386	83.6062
600	360 000	24.4949	77.4597	650	422 500	25.4951	80.6226	700	490 000	26.4575	83.6660

N	N²	√N̄	√1̄0̄N̄	N	N²	√N̄	√1̄0̄N̄	N	N²	√N̄	√1̄0̄N̄
700	490 000	26.4575	83.6660	750	562 500	27.3861	86.6025	800	640 000	28.2843	89.4427
701	491 401	26.4764	83.7257	751	564 001	27.4044	86.6603	801	641 601	28.3019	89.4986
702	492 804	26.4953	83.7854	752	565 504	27.4226	86.7179	802	643 204	28.3196	89.5545
703	494 209	26.5141	83.8451	753	567 009	27.4408	86.7756	803	644 809	28.3373	89.6103
704	495 616	26.5350	83.9047	754	568 516	27.4591	86.8332	804	646 416	28.3549	89.6660
705	497 025	26.5518	83.9643	755	570 025	27.4773	86.8907	805	648 025	28.3725	89.7218
706	498 436	26.5707	84.0238	756	571 536	27.4955	86.9483	806	649 636	28.3901	89.7775
707	499 849	26.5895	84.0833	757	573 049	27.5136	87.0057	807	651 249	28.4077	89.8332
708	501 264	26.6083	84.1427	758	574 564	27.5318	87.0632	808	652 864	28.4253	89.8888
709	502 681	26.6271	84.2021	759	576 081	27.5500	87.1206	809	654 481	28.4429	89.9444
710	504 100	26.6458	84.2615	760	577 600	27.5681	87.1780	810	656 100	28.4605	90.0000
711	505 521	26.6646	84.3208	761	579 121	27.5862	87.2353	811	657 721	28.4781	90.0555
712	506 944	26.6833	84.3801	762	580 644	27.6043	87.2926	812	659 344	28.4956	90.1110
713	508 369	26.7021	84.4393	763	582 169	27.6225	87.3499	813	660 969	28.5132	90.1665
714	509 796	26.7208	84.4985	764	583 696	27.6405	87.4071	814	662 596	28.5307	90.2219
715	511 225	26.7395	84.5577	765	585 225	27.6586	87.4643	815	664 225	28.5482	90.2774
716	512 656	26.7582	84.6168	766	586 756	27.6767	87.5214	816	665 856	28.5657	90.3327
717	514 089	26.7769	84.6759	767	588 289	27.6948	87.5785	817	667 489	28.5832	90.3881
718	515 524	26.7955	84.7349	768	589 824	27.7128	87.6356	818	669 124	28.6007	90.4434
719	516 961	26.8142	84.7939	769	591 361	27.7308	87.6926	819	670 761	28.6182	90.4986
720	518 400	26.8328	84.8528	770	592 900	27.7489	87.7496	820	672 400	28.6356	90.5539
721	519 841	26.8514	84.9117	771	594 441	27.7669	87.8066	821	674 041	28.6531	90.6091
722	521 284	26.8701	84.9706	772	595 984	27.7849	87.8635	822	675 684	28.6705	90.6642
723	522 729	26.8887	85.0294	773	597 529	27.8029	87.9204	823	677 329	28.6880	90.7193
724	524 176	26.9072	85.0882	774	599 076	27.8209	87.9773	824	678 976	28.7054	90.7744
725	525 625	26.9258	85.1469	775	600 625	27.8388	88.0341	825	680 625	28.7228	90.8295
726	527 076	26.9444	85.2056	776	602 176	27.8568	88.0909	826	682 276	28.7402	90.8845
727	528 529	26.9629	85.2643	777	603 729	27.8747	88.1476	827	683 929	28.7576	90.9395
728	529 984	26.9815	85.3229	778	605 284	27.8927	88.2043	828	685 584	28.7750	90.9945
729	531 441	27.0000	85.3815	779	606 841	27.9106	88.2610	829	687 241	28.7924	91.0494
730	532 900	27.0185	85.4400	780	608 400	27.9285	88.3176	830	688 900	28.8097	91.1043
731	534 361	27.0370	85.4985	781	609 961	27.9464	88.3742	831	690 561	28.8271	91.1592
732	535 824	27.0555	85.5570	782	611 524	27.9643	88.4308	832	692 224	28.8444	91.2140
733	537 289	27.0740	85.6154	783	613 089	27.9821	88.4873	833	693 889	28.8617	91.2688
734	538 756	27.0924	85.6738	784	614 656	28.0000	88.5438	834	695 556	28.8791	91.3236
735	540 225	27.1109	85.7321	785	616 225	28.0179	88.6002	835	697 225	28.8964	91.3783
736	541 696	27.1293	85.7904	786	617 796	28.0357	88.6566	836	698 896	28.9137	91.4330
737	543 169	27.1477	85.8487	787	619 369	28.0535	88.7130	837	700 569	28.9310	91.4877
738	544 644	27.1662	85.9069	788	620 944	28.0713	88.7694	838	702 244	28.9482	91.5423
739	546 121	27.1846	85.9651	789	622 521	28.0891	88.8257	839	703 921	28.9655	91.5969
740	547 600	27.2029	86.0233	790	624 100	28.1069	88.8819	840	705 600	28.9828	91.6515
741	549 081	27.2213	86.0814	791	625 681	28.1247	88.9382	841	707 281	29.0000	91.7061
742	550 564	27.2397	86.1394	792	627 264	28.1425	88.9944	842	708 964	29.0172	91.7606
743	552 049	27.2580	86.1974	793	628 849	28.1603	89.0505	843	710 649	29.0345	91.8150
744	553 536	27.2764	86.2554	794	630 436	28.1780	89.1067	844	712 336	29.0517	91.8695
745	555 025	27.2947	86.3134	795	632 025	28.1957	89.1628	845	714 025	29.0689	91.9239
746	556 516	27.3130	86.3713	796	633 616	28.2135	89.2188	846	715 716	29.0861	91.9783
747	558 009	27.3313	86.4292	797	635 209	28.2312	89.2749	847	717 409	29.1033	92.0326
748	559 504	27.3496	86.4870	798	636 804	28.2489	89.3308	848	719 104	29.1204	92.0869
749	561 001	27.3679	86.5448	799	638 401	28.2666	89.3868	849	720 801	29.1376	92.1412
750	562 500	27.3861	86.6025	800	640 000	28.2843	89.4427	850	722 500	29.1548	92.1954

N	N²	√N	√10N	N	N²	√N	√10N	N	N²	√N	√10N
850	722 500	29.1548	92.1954	900	810 000	30.0000	94.8683	950	902 500	30.8221	97.4679
851	724 201	29.1719	92.2497	901	811 810	30.0167	94.9210	951	904 401	30.8383	97.5192
852	725 904	29.1890	92.3038	902	813 604	30.0333	94.9737	952	906 304	30.8545	97.5705
853	727 609	29.2062	92.3580	903	815 409	30.0500	95.0263	953	908 209	30.8707	97.6217
854	729 316	29.2233	92.4121	904	817 216	30.0666	95.0789	954	910 116	30.8869	97.6729
855	731 025	29.2404	92.4662	905	819 025	30.0832	95.1315	955	912 025	30.9031	97.7241
856	732 736	29.2575	92.5203	906	820 836	30.0998	95.1840	956	913 936	30.9192	97.7753
857	734 449	29.2746	92.5743	907	822 649	30.1164	95.2365	957	915 849	30.9354	97.8264
858	736 164	29.2916	92.6283	908	824 464	30.1330	95.2890	958	917 764	30.9516	97.8775
859	737 881	29.3087	92.6823	909	826 281	30.1496	95.3415	959	919 681	30.9677	97.9285
860	739 600	29.3258	92.7362	910	828 100	30.1662	95.3939	960	921 600	30.9839	97.9796
861	741 321	29.3428	92.7901	911	829 921	30.1828	95.4463	961	923 521	31.0000	98.0306
862	743 044	29.3598	92.8440	912	831 744	30.1993	95.4987	962	925 444	31.0161	98.0816
863	744 769	29.3769	92.8978	913	833 569	30.2159	95.5510	963	927 369	31.0322	98.1326
864	746 496	29.3939	92.9516	914	835 396	30.2324	95.6033	964	929 296	31.0483	98.1835
865	748 225	29.4109	93.0054	915	837 225	30.2490	95.6556	965	931 225	31.0644	98.2344
866	749 956	29.4279	93.0591	916	839 056	30.2655	95.7079	966	933 156	31.0805	98.2853
867	751 689	29.4449	93.1128	917	840 889	30.2820	95.7601	967	935 089	31.0966	98.3362
868	753 424	29.4618	93.1665	918	842 724	30.2985	95.8123	968	937 024	31.1127	98.3870
869	755 161	29.4788	93.2202	919	844 561	30.3150	95.8645	969	938 961	31.1288	98.4378
870	756 900	29.4958	93.2738	920	846 400	30.3315	95.9166	970	940 900	31.1448	98.4886
871	758 641	29.5127	93.3274	921	848 241	30.3480	95.9687	971	942 841	31.1609	98.5393
872	760 384	29.5296	93.3809	922	850 084	30.3645	96.0208	972	944 784	31.1769	98.5901
873	762 129	29.5466	93.4345	923	851 929	30.3809	96.0729	973	946 729	31.1929	98.6408
874	763 876	29.5635	93.4880	924	853 776	30.3974	96.1249	974	948 676	31.2090	98.6914
875	765 625	29.5804	93.5414	925	855 625	30.4138	96.1769	975	950 625	31.2250	98.7421
876	767 376	29.5973	93.5949	926	857 476	30.4302	96.2289	976	952 576	31.2410	98.7927
877	769 129	29.6142	93.6483	927	859 329	30.4467	96.2808	977	954 529	31.2570	98.8433
878	770 884	29.6311	93.7017	928	861 184	30.4631	96.3329	978	956 484	31.2730	98.8939
879	772 641	29.6479	93.7550	929	863 041	30.4795	96.3846	979	958 441	31.2890	98.9444
880	774 400	29.6648	93.8083	930	864 900	30.4959	96.4365	980	960 400	31.3050	98.9949
881	776 161	29.6816	93.8616	931	866 761	30.5123	96.4883	981	962 361	31.3209	99.0454
882	777 924	29.6985	93.9149	932	868 624	30.5287	96.5401	982	964 324	31.3369	99.0959
883	779 689	29.7153	93.9681	933	870 489	30.5450	96.5919	983	966 289	31.3528	99.1464
884	781 456	29.7321	94.0213	934	872 356	30.5614	96.6437	984	968 256	31.3688	99.1968
885	783 225	29.7489	94.0744	935	874 225	30.5778	96.6954	985	970 225	31.3847	99.2472
886	784 996	29.7658	94.1276	936	876 096	30.5941	96.7471	986	972 196	31.4006	99.2975
887	786 769	29.7825	94.1807	937	877 969	30.6105	96.7988	987	974 169	31.4116	99.3479
888	788 544	29.7993	94.2338	938	879 844	30.6268	96.8504	988	976 144	31.4325	99.3982
889	790 321	29.8161	94.2868	939	881 721	30.6431	96.9020	989	978 121	31.4484	99.4485
890	792 100	29.8329	94.3398	940	883 600	30.6594	96.9536	990	980 100	31.4643	99.4987
891	793 881	29.8496	94.3928	941	885 481	30.6757	97.0052	991	982 081	31.4802	99.5490
892	795 664	29.8664	94.4458	942	887 364	30.6920	97.0567	992	984 064	31.4960	99.5992
893	797 449	29.8831	94.4987	943	889 249	30.7083	97.1082	993	986 049	31.5119	99.6494
894	799 236	29.8998	94.5516	944	891 136	30.7246	97.1597	994	988 036	31.5278	99.6995
895	801 025	29.9166	94.6044	945	893 025	30.7409	97.2111	995	990 025	32.5436	99.7497
896	802 816	29.9333	94.6573	946	894.916	30.7571	97.2625	996	992 016	31.5595	99.7998
897	804 609	29.9500	94.7107	947	896 809	30.7734	97.3139	997	994 009	31.5753	99.8499
898	806 404	29.9666	94.7629	948	898 704	30.7896	97.3653	998	996 004	31.5911	99.8999
899	808 201	29.9833	94.8156	949	900 601	30.8058	97.4166	999	998 001	31.6070	99.9500
900	810 000	30.0000	94.8683	950	902 500	30.8221	97.4679	1000	1000 000	31.6228	100.0000

TABLE A-2. The Binomial Distribution
For designated values of n and p, the tabled entry gives $P(Y \leq y)$.

n	y	$p = .10$	$p = .20$	$p = .25$	$p = .30$	$p = .40$	$p = .50$
5	0	.5905	.3277	.2373	.1681	.0778	.0312
	1	.9185	.7373	.6328	.5282	.3370	.1875
	2	.9914	.9421	.8965	.8369	.6826	.5000
	3	.9995	.9933	.9844	.9692	.9130	.8125
	4	1.0000	.9997	.9990	.9976	.9898	.9688
	5		1.0000	1.0000	1.0000	1.0000	1.0000
10	0	.3487	.1074	.0563	.0282	.0060	.0010
	1	.7361	.3758	.2440	.1493	.0464	.0107
	2	.9298	.6778	.5256	.3828	.1673	.0547
	3	.9872	.8791	.7759	.6496	.3823	.1719
	4	.9984	.9672	.9219	.8497	.6331	.3770
	5	.9999	.9936	.9803	.9527	.8338	.6230
	6	1.0000	.9991	.9965	.9894	.9452	.8281
	7		.9999	.9996	.9984	.9877	.9453
	8		1.0000	1.0000	.9999	.9983	.9893
	9				1.0000	.9999	.9990
	10					1.0000	1.0000
15	0	.2059	.0352	.0134	.0047	.0005	.0000
	1	.5490	.1671	.0802	.0353	.0052	.0005
	2	.8159	.3980	.2361	.1268	.0271	.0037
	3	.9444	.6482	.4613	.2969	.0905	.0176
	4	.9873	.8358	.6865	.5155	.2173	.0592
	5	.9978	.9389	.8516	.7216	.4032	.1509
	6	.9997	.9819	.9434	.8689	.6098	.3036
	7	1.0000	.9958	.9827	.9500	.7869	.5000
	8		.9992	.9958	.9848	.9050	.6964
	9		.9999	.9992	.9963	.9662	.8491
	10		1.0000	.9999	.9993	.9907	.9408
	11			1.0000	.9999	.9981	.9824
	12				1.0000	.9997	.9963
	13					1.0000	.9995
	14						1.0000
20	0	.1216	.0115	.0032	.0008	.0000	.0000
	1	.3917	.0692	.0243	.0076	.0005	.0000
	2	.6769	.2061	.0913	.0355	.0036	.0002
	3	.8670	.4114	.2252	.1071	.0160	.0013
	4	.9568	.6296	.4148	.2375	.0510	.0059
	5	.9887	.8042	.6172	.4164	.1256	.0207
	6	.9976	.9133	.7858	.6080	.2500	.0577
	7	.9996	.9679	.8982	.7723	.4159	.1316
	8	.9999	.9900	.9591	.8867	.5956	.2517
	9	1.0000	.9974	.9861	.9520	.7553	.4119
	10		.9994	.9961	.9829	.8725	.5881
	11		.9999	.9991	.9949	.9435	.7483
	12		1.0000	.9998	.9987	.9790	.8684
	13			1.0000	.9997	.9935	.9423
	14				1.0000	.9984	.9793
	15					.9997	.9941
	16					1.0000	.9987
	17						.9998
	18						1.0000

TABLE A-3. The Standard Normal Distribution

z	$P(Z \leq z)$	z	$P(Z \leq z)$	z	$P(Z \leq z)$	z	$P(Z \leq z)$
−4.265	.00001	−1.6	.0548	0	.5000	1.645	.9500
−4.0	.00003	−1.555	.0600			1.7	.9554
		−1.5	.0668	0.1	.5398	1.751	.9600
−3.8	.00007			0.126	.5500	1.8	.9641
−3.719	.0001	−1.476	.0700	0.2	.5793	1.881	.9700
−3.6	.0002	−1.405	.0800	0.253	.6000	1.9	.9713
		−1.4	.0808	0.3	.6179	1.960	.9750
−3.4	.0003	−1.341	.0900	0.385	.6500		
−3.291	.0005	−1.3	.0968	0.4	.6554	2.0	.9773
−3.2	.0007	−1.282	.1000			2.054	.9800
−3.090	.0010	−1.2	.1151	0.5	.6915	2.1	.9821
−3.0	.0013	−1.1	.1357	0.524	.7000	2.2	.9861
		−1.036	.1500	0.6	.7258	2.3	.9893
−2.9	.0019	−1.0	.1587	0.674	.7500	2.326	.9900
−2.8	.0026			0.7	.7580	2.4	.9918
−2.7	.0035	−0.9	.1841	0.8	.7881		
−2.6	.0047	−0.842	.2000	0.842	.8000	2.5	.9938
−2.576	.0050	−0.8	.2119	0.9	.8159	2.576	.9950
−2.5	.0062	−0.7	.2420			2.6	.9953
		−0.674	.2500	1.0	.8413	2.7	.9965
−2.4	.0082	−0.6	.2742	1.036	.8500	2.8	.9974
−2.326	.0100	−0.524	.3000	1.1	.8643	2.9	.9981
−2.3	.0107	−0.5	.3085	1.2	.8849		
−2.2	.0139			1.282	.9000	3.0	.9987
−2.1	.0179	−0.4	.3446	1.3	.9032	3.090	.9990
−2.054	.0200	−0.385	.3500	1.341	.9100	3.2	.9993
−2.0	.0227	−0.3	.3821	1.4	.9192	3.291	.9995
		−0.253	.4000	1.405	.9200	3.4	.9997
−1.960	.0250	−0.2	.4207	1.476	.9300		
−1.9	.0287	−0.126	.4500			3.6	.9998
−1.881	.0300	−0.1	.4602	1.5	.9332	3.719	.9999
−1.8	.0359			1.555	.9400	3.8	.99993
−1.751	.0400	0	.5000	1.6	.9452		
−1.7	.0446					4.0	.99997
−1.645	.0500					4.265	.99999

TABLE A-4. Percentiles of the t Distributions

d.f. \ %	55	65	75	85	90	95	97.5	99	99.5	99.95
1	0.158	0.510	1.000	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.445	0.816	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.137	0.424	0.765	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.414	0.741	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.408	0.727	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.404	0.718	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.402	0.711	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.399	0.706	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.398	0.703	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.397	0.700	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.396	0.697	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.128	0.395	0.695	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.128	0.394	0.694	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.128	0.393	0.692	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.128	0.393	0.691	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.128	0.392	0.690	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.128	0.392	0.689	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.127	0.392	0.688	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.127	0.391	0.688	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.127	0.391	0.687	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.127	0.391	0.686	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.127	0.390	0.686	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.127	0.390	0.685	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.127	0.390	0.685	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.127	0.390	0.684	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.127	0.390	0.684	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.127	0.389	0.684	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.127	0.389	0.683	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.127	0.389	0.683	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.127	0.389	0.683	1.055	1.310	1.697	2.042	2.457	2.750	3.646
35	0.127	0.388	0.682	1.052	1.306	1.690	2.030	2.438	2.724	3.591
40	0.126	0.388	0.681	1.050	1.303	1.684	2.021	2.423	2.704	3.551
45	0.126	0.388	0.680	1.049	1.301	1.679	2.014	2.412	2.690	3.520
50	0.126	0.388	0.679	1.047	1.299	1.676	2.009	2.403	2.678	3.496
60	0.126	0.387	0.679	1.045	1.296	1.671	2.000	2.390	2.660	3.460
70	0.126	0.387	0.678	1.044	1.294	1.667	1.994	2.381	2.648	3.435
80	0.126	0.387	0.678	1.043	1.292	1.664	1.990	2.374	2.639	3.416
90	0.126	0.387	0.677	1.042	1.291	1.662	1.987	2.368	2.632	3.402
100	0.126	0.386	0.677	1.042	1.290	1.660	1.984	2.364	2.626	3.390
120	0.126	0.386	0.677	1.041	1.289	1.658	1.980	2.358	2.617	3.373
140	0.126	0.386	0.676	1.040	1.288	1.656	1.977	2.353	2.611	3.361
160	0.126	0.386	0.676	1.040	1.287	1.654	1.975	2.350	2.607	3.352
180	0.126	0.386	0.676	1.039	1.286	1.653	1.973	2.347	2.603	3.345
200	0.126	0.386	0.676	1.039	1.286	1.653	1.972	2.345	2.601	3.340
∞	0.126	0.385	0.674	1.036	1.282	1.645	1.960	2.326	2.576	3.291

% d.f.	0.5	1	2.5	5	10	20	30	40	50
1	0.00004	0.0002	0.001	0.004	0.016	0.064	0.148	0.275	0.455
2	0.010	0.020	0.051	0.103	0.211	0.446	0.713	1.022	1.386
3	0.072	0.115	0.216	0.352	0.584	1.005	1.424	1.869	2.366
4	0.207	0.297	0.484	0.711	1.064	1.649	2.195	2.753	3.357
5	0.412	0.554	0.831	1.145	1.610	2.343	3.000	3.655	4.351
6	0.676	0.872	1.237	1.635	2.204	3.070	3.828	4.570	5.348
7	0.989	1.239	1.690	2.167	2.833	3.822	4.671	5.493	6.346
8	1.344	1.646	2.180	2.733	3.490	4.594	5.527	6.423	7.344
9	1.735	2.088	2.700	3.325	4.168	5.380	6.393	7.357	8.343
10	2.156	2.558	3.247	3.940	4.865	6.179	7.267	8.295	9.342
11	2.603	3.053	3.816	4.575	5.578	6.989	8.148	9.237	10.341
12	3.074	3.571	4.404	5.226	6.304	7.807	9.034	10.182	11.340
13	3.565	4.107	5.009	5.892	7.042	8.634	9.926	11.129	12.340
14	4.075	4.660	5.629	6.571	7.790	9.467	10.821	12.078	13.339
15	4.601	5.229	6.262	7.261	8.547	10.307	11.721	13.030	14.339
16	5.142	5.812	6.908	7.962	9.312	11.152	12.624	13.983	15.338
17	5.697	6.408	7.564	8.672	10.085	12.002	13.531	14.937	16.338
18	6.265	7.015	8.231	9.390	10.865	12.857	14.440	15.893	17.338
19	6.844	7.633	8.907	10.117	11.651	13.716	15.352	16.850	18.338
20	7.434	8.260	9.591	10.851	12.443	14.578	16.266	17.809	19.337
21	8.034	8.897	10.283	11.591	13.240	15.445	17.182	18.768	20.337
22	8.643	9.542	10.982	12.338	14.041	16.314	18.101	19.729	21.337
23	9.260	10.196	11.689	13.091	14.848	17.187	19.021	20.690	22.337
24	9.886	10.856	12.401	13.848	15.659	18.062	19.943	21.752	23.337
25	10.520	11.524	13.120	14.611	16.473	18.940	20.867	22.616	24.337
26	11.160	12.198	13.844	15.379	17.292	19.820	21.792	23.579	25.336
27	11.808	12.879	14.573	16.151	18.114	20.703	22.719	24.544	26.336
28	12.461	13.565	15.308	16.928	18.939	21.588	23.647	25.509	27.336
29	13.121	14.256	16.047	17.708	19.768	22.475	24.577	26.475	28.336
30	13.787	14.953	16.791	18.493	20.599	23.364	25.508	27.442	29.336
35	17.192	18.509	20.569	22.465	24.797	27.836	30.178	32.282	34.336
40	20.707	22.164	24.433	26.509	29.051	32.345	34.872	37.134	39.335
45	24.311	25.901	28.366	30.612	33.350	36.884	39.585	41.995	44.335
50	27.991	29.707	32.357	34.764	37.689	41.449	44.313	46.864	49.335
60	35.534	37.485	40.482	43.188	46.459	50.641	53.809	56.620	59.335
70	43.275	45.442	48.758	51.739	55.329	59.898	63.346	66.396	69.334
80	51.172	53.540	57.153	60.391	64.278	69.207	72.915	76.188	79.334
90	59.196	61.754	65.647	69.126	73.291	78.558	82.511	85.993	89.334
100	67.328	70.065	74.222	77.929	82.358	87.945	92.129	95.808	99.334
120	83.852	86.923	91.573	95.705	100.624	106.806	111.419	115.465	119.334
140	100.655	104.034	109.137	113.659	119.029	125.758	130.766	135.149	139.334
160	117.679	121.346	126.870	131.756	137.546	144.783	150.158	154.856	159.334
180	134.884	138.820	144.741	149.969	156.153	163.868	169.588	174.580	179.334
200	152.241	156.432	162.728	168.279	174.835	183.003	189.049	194.319	199.334

χ^2 Distributions

60	70	80	90	95	97.5	99	99.5	99.95	% / d.f.
0.708	1.074	1.642	2.706	3.841	5.024	6.635	7.879	12.116	1
1.833	2.408	3.219	4.605	5.991	7.378	9.210	10.597	15.202	2
2.946	3.665	4.642	6.251	7.815	9.348	11.345	12.838	17.730	3
4.045	4.878	5.989	7.779	9.488	11.143	13.277	14.860	19.997	4
5.132	6.064	7.289	9.236	11.070	12.833	15.086	16.750	22.105	5
6.211	7.231	8.558	10.645	12.592	14.449	16.812	18.548	24.103	6
7.283	8.383	9.803	12.017	14.067	16.013	18.475	20.278	26.018	7
8.351	9.524	11.030	13.362	15.507	17.535	20.090	21.955	27.868	8
9.414	10.656	12.242	14.684	16.919	19.023	21.666	23.589	29.666	9
10.473	11.781	13.442	15.987	18.307	20.483	23.209	25.188	31.420	10
11.530	12.899	14.631	17.275	19.675	21.920	24.725	26.757	33.137	11
12.584	14.011	15.812	18.549	21.026	23.337	26.217	28.300	34.821	12
13.636	15.119	16.985	19.812	22.362	24.736	27.688	29.819	36.478	13
14.685	16.222	18.151	21.064	23.685	26.119	29.141	31.319	38.109	14
15.733	17.322	19.311	22.307	24.996	27.488	30.578	32.801	39.719	15
16.780	18.418	20.465	23.542	26.296	28.845	32.000	34.267	41.308	16
17.824	19.511	21.615	24.769	27.587	30.191	33.409	35.718	42.879	17
18.868	20.601	22.760	25.989	28.869	31.526	34.805	37.156	44.434	18
19.910	21.689	23.900	27.204	30.144	32.852	36.191	38.582	45.973	19
20.951	22.775	25.038	28.412	31.410	34.170	37.566	39.997	47.498	20
21.991	23.858	26.171	29.615	32.671	35.479	38.932	41.401	49.011	21
23.031	24.939	27.301	30.813	33.924	36.781	40.289	42.796	50.511	22
24.069	26.018	28.429	32.007	35.172	38.076	41.638	44.181	52.000	23
25.106	27.096	29.553	33.196	36.415	39.364	42.980	45.559	53.479	24
26.143	28.172	30.675	34.382	37.652	40.646	44.314	46.928	54.947	25
27.179	29.246	31.795	35.563	38.885	41.923	45.642	48.290	56.407	26
28.214	30.319	32.912	36.741	40.113	43.195	46.963	49.645	57.858	27
29.249	31.391	34.027	37.916	41.337	44.461	48.278	50.993	59.300	28
30.283	32.461	35.139	39.087	42.557	45.722	49.588	52.336	60.735	29
31.316	33.530	36.250	40.256	43.773	46.979	50.892	53.672	62.162	30
36.475	38.859	41.778	46.059	49.802	53.203	57.342	60.275	69.199	35
41.622	44.165	47.269	51.805	55.758	59.342	63.691	66.766	76.095	40
46.761	49.452	52.729	57.505	61.656	65.410	69.957	73.166	82.876	45
51.892	54.723	58.164	63.167	67.505	71.420	76.154	79.490	89.561	50
62.135	65.227	68.972	74.397	79.082	83.298	88.379	91.952	102.695	60
72.358	75.689	79.715	85.527	90.531	95.023	100.425	104.215	115.578	70
82.566	86.120	90.405	96.578	101.879	106.629	112.329	116.321	128.261	80
92.761	96.524	101.054	107.565	113.145	118.136	124.116	128.299	140.782	90
102.946	106.906	111.667	118.498	124.342	129.561	135.807	140.169	153.167	100
123.289	127.616	132.806	140.233	146.567	152.211	158.950	163.648	177.603	120
143.604	148.269	153.854	161.827	168.613	174.648	181.840	186.847	201.683	140
163.898	168.876	174.828	183.311	190.516	196.915	204.530	209.824	225.481	160
184.173	189.446	195.743	204.704	212.304	219.044	227.056	232.620	249.048	180
204.434	209.985	216.609	226.021	233.994	241.058	249.445	255.264	272.423	200

TABLE A-6 A Short Table of Random Digits[a]

03000	95429	05023	42445	81479	06582	57832	19864	09655	46091	39898
03001	54933	30376	60217	12916	71034	41493	00810	40746	22328	98099
03002	49242	61815	36878	72093	63479	15847	81812	72644	85935	82053
03003	41240	22056	66879	48908	44909	12820	75666	70382	86600	21025
03004	02049	38223	10899	13677	14360	74016	06527	80281	53788	70695
03005	27000	12588	72677	50662	96047	51209	61781	22706	94834	24773
03006	91285	54345	34034	37310	55291	63399	98036	17711	15957	26772
03007	44887	26995	74237	88921	18037	92664	90519	13201	54268	92950
03008	72892	36748	42544	00765	29826	04582	16897	05507	00115	95513
03009	17812	43757	07029	78410	93762	09606	33152	81105	79698	22892
03010	95921	10413	86215	85039	35246	27026	13873	35350	94513	38339
03011	40023	28286	92943	41583	27563	73009	52091	86401	64081	10484
03012	80560	81722	72870	86454	57429	72880	60952	61152	02839	30079
03013	47712	67899	21900	18132	29785	26865	42085	97353	43889	41507
03014	80406	13822	81956	08991	95359	03425	39800	53545	16848	26169
03015	56642	53730	56710	70241	10522	30170	47951	28314	84072	86847
03016	74682	44393	69167	41446	43982	45485	23758	88922	76378	76329
03017	45524	46409	12928	46822	21407	04653	03029	95903	71550	87652
03018	57139	42113	85637	92492	06469	35989	92453	68124	98263	75567
03019	00705	17743	58687	88502	80485	22352	93014	94452	18065	42134
03020	05549	70101	92945	27189	95527	49801	05829	59762	74343	51864
03021	22327	55480	95907	77989	88732	93567	75074	32414	81044	77751
03022	86018	85335	88367	10506	33585	31568	93165	63832	06743	65506
03023	70839	18951	33484	00091	74367	72892	04084	59770	82641	32061
03024	67411	88645	77789	72757	59927	39876	87841	51595	96364	60506
03025	28259	11822	53058	16516	20196	71098	63699	63075	68394	32666
03026	04503	79028	63404	03482	76677	23288	39594	66119	81274	99982
03027	63288	93134	91378	91988	12157	34581	79674	85790	22454	87109
03028	63701	54429	86992	02004	28351	13700	89012	70678	92686	89488
03029	96540	71159	93478	97919	43501	65527	82167	15386	33527	12041
03030	16744	20236	89506	63437	81051	56704	59752	67516	93433	82926
03031	75423	86481	84147	82941	47088	75897	62126	14258	15839	82679
03032	48129	76751	53494	04187	79955	30747	81231	26399	40379	16182
03033	81433	53248	63646	64467	48089	89590	15504	16287	02513	03315
03034	81800	27936	60474	34712	90626	62663	68216	39411	28219	98681
03035	07751	33731	44285	96913	33077	93006	52025	19420	77286	92896
03036	26480	31002	93898	49074	16354	28824	09677	57532	14898	43899
03037	09458	04923	17904	70053	69033	81766	54408	03839	04008	69508
03038	91478	94042	90945	63299	86572	67303	55861	78772	46641	20170
03039	36033	56105	41641	23328	59888	47185	72830	71950	59564	97746
03040	22404	60701	89749	20791	60898	06962	75470	23908	18351	08554
03041	94658	81609	05446	72224	93683	52224	17750	48138	41620	90644
03042	39921	05493	34544	93414	02831	02695	06558	13437	91279	30904
03043	36744	75178	91131	27882	87981	82359	62162	72337	98921	74546
03044	16956	15978	46143	32638	46917	23087	69066	99985	01735	20241
03045	33291	76222	52088	07482	83308	98960	40993	84206	76311	91982
03046	63079	72605	89507	42852	14681	18795	78102	30630	16175	77998
03047	47767	52203	62663	32254	31966	15805	99528	34002	69239	77643
03048	76432	63190	29948	04588	27277	11293	09048	55210	71579	95191
03049	04180	18717	58573	32422	38615	85666	03020	27931	01660	44950

[a] Reprinted by permission from RAND Corporation, *A Million Random Digits*, The Free Press (1955), p. 61.

Appendix B
NUMERICAL ANSWERS

2.2.2. Cell entries are:

33	27		49	14	8	78	31
29	21						
5	1						
37	27						

2.2.5. a. 37, 8; b. 32 liked, 5 did not, 8 had no opinion; c. 86.5%.

2.2.7. Cell entries are:

10	14	1	0	25
8	1	0	0	9
10	1	0	0	11
28	16	1	0	45

2.2.8. Cell entries are:

3	4	9	0	16
0	4	10	3	17
4	5	1	0	10
1	0	1	0	2
8	13	21	3	45

a. 10; b. 3, 6.67%; c. 82.22%; d. 64.44%.

2.3.1 Political party preference	Opinion on legalizing marijuana					
	Strongly disagree	Mildly disagree	No opinion	Mildly agree	Strongly agree	Totals
Democrat	19	3	2	26	10	60
Republican	17	7	3	22	1	50
Other party	1	1	1	3	0	6
No preference	12	3	2	27	20	64
Totals	49	14	8	78	31	180

2.3.5. c. (a) 67, (b) 66, (c) 10–19.

3.2.1. a. 423.6; b. 423; c. 426; d. 424.

3.2.2. a. mean = 11.

3.2.3. a. 36.20; b. 36.3; c. there is no single mode.

3.2.4. a. mean = 102.9, median = 104, midrange = 104.
b. 60%. c. 15%. d. 102,900; 98,760.

3.2.5. a. mean = median = midrange = 1.2 ounces.

3.2.6. a. 50.2. b. there is no mode; median = 45.

3.2.7.

	Volume	Last price	Net change
Mean	720,180	31.367	+1.133
Median	718,500	$28\frac{1}{4}$	$+1\frac{1}{4}$

3.2.8. i. 43%; ii. $\frac{2}{3}$ are H; iii. 16.5; iv. 47.4.

3.2.9. a. line #1: 1.213, 1.22, 1.22; line #2: 1.244, 1.22, 1.22.
c. 1.229.

3.2.10. a. 12,518.2; 10,700; 10,500; 13,750.

3.2.11. a. 14,443; b. mean = 1444.3, median = 1714.5.

3.2.13. a. 9,025.

3.2.15. a. mean = 201 (201.2).

3.2.16. c. 183.4; d. 183.

3.2.17. a. 62,550; b. 54,000; e. roughly 45,000.

3.2.20. (a) 88.4; (b) 88; (c) 88.

3.2.21. (a) 67.02; (b) 67; (c) 67.

3.2.22. a. 70.7, 71, 70.

3.2.23. b. 20, 26.25, 20.75, 20.5, 19.5, 18.75.

3.4.1. 12, 11.94, 3.45, 0.81%.

3.4.2. 8, 5.64, 2.37, 21.5%.

3.4.3. 2.5, 0.375, 0.612, 1.69%.

3.4.4. 0.8, 0.0667, 0.258, 21.5%.

3.4.5. Line #1: 0.16, 0.00283, 0.0532, 4.39%.
Line #2: 0.14, 0.00223, 0.0472, 3.79%.

3.4.6. 58, 316.37, 17.79, 28.44%.

3.4.7. 26, 63.41, 7.96, 9.00%.

3.4.8. 20, 6.62, 2.57, 3.83%.

3.4.9. 37, 140.9, 11.87, 16.79%.

3.5.1. b. $5,000; $2,108.21; $2,058.51; $854.80; 40.55%.

3.5.2.

	Sample #1	Sample #2	Sample #3	Sample #4
a. (i) male	71.11%	51.11%	62.22%	46.67%
(ii) female	28.89%	48.89%	37.78%	53.33%
b. mean	15.60	11.48	12.52	10.49
median	10.00	11.00	8.00	7.50
s.d.	12.48	10.72	10.95	8.79
c.v.	80.00%	93.38%	87.46%	83.79%

d. Percentage agreeing to legalization:

	Sample #1	Sample #2	Sample #3	Sample #4
Overall	51.11	62.22	71.11	55.56
Male	43.75	69.57	85.71	57.14
Female	69.23	54.55	47.06	54.17
Democrat	58.82	71.43	53.33	50.00
Republican	33.33	18.18	83.33	53.85
Other party	0.00	—	50.00	100.00
No party	72.73	80.00	81.25	56.25

e. Grade point average (G.P.A.):

	Sample #1	Sample #2	Sample #3	Sample #4
Mean	2.39	2.48	2.43	2.31
Standard deviation	0.65	0.48	0.62	0.51
Mean: male	2.35	2.46	2.42	2.24
S.d.: male	0.74	0.56	0.63	0.50
Mean: female	2.50	2.50	2.43	2.37
S.d.: female	0.36	0.39	0.61	0.51
Mean: 0–9.5 mi.	2.07	2.35	2.55	2.29
10–19.5 mi.	2.68	2.64	2.28	2.49
20–29.5 mi.	2.48	2.40	2.38	2.16
30–39.5 mi.	2.93	—	2.14	2.37
40–49.5 mi.	—	—	2.08	—
50–59.5 mi.	3.31	3.37	—	—

4.3.1. 1/8, 3/8, 1/8.

4.3.2. P(0 head) = 1/16   P(3 heads) = 1/4,
        P(1 head) = 1/4,   P(4 heads) = 1/16.
        P(2 heads) = 3/8,

4.3.3. 1/36, 1/18, 1/12, 1/9, 5/36, 1/6, 5/36, 1/9, 1/12, 1/18, 1/36.

4.3.4. (a) 1/4, (b) 1/2, (c) 1/13, (d) 3/13, (e) 8/13 [9/13 if *ace* is counted as "one"], (f) 1/26.

4.3.5. (a) 19/45, (b) 5/18, (c) 2/45, (d) 7/30.

4.3.6. 21/76, 1/38, 15/76.

4.3.7. 31/104, 8/19.

4.3.8. 12/53, 1/2.

4.3.9. 1/3, 7/12, 2/7.

4.3.10. 2/3, 2/5.

4.3.11. 1/3, 10/27, 8/27.

4.5.1. (a) 1/1024, (b) 33/66640.

4.5.2. (a) 1/256, (b) 33/16660.

4.5.3. No pair shows independence: 29/180 vs. 13/81, 1/36 vs. 13/675, 37/180 vs. 416/2025, 3/20 vs. 19/135, 7/60 vs. 19/162, 1/180 vs. 19/1350, 3/20 vs. 304/2025.

4.5.4. No: 1/90 vs. 32/2025.

4.5.5. .2277.

4.5.6. a. 0.4096, b. 0.6723, c. 0.5886.

4.5.7. .0905.

4.5.8. (a) .0139, (b) .0000 (less than .00005).

4.5.9. .6778.

4.5.10. 10.

4.7.1. (a) .2119   (e) .8849
        (b) .9773   (f) .1151
        (c) .1587   (g) .1151
        (d) .5793   (h) .8849

4.7.2. (a) .6710   (d) .8664
        (b) .0501   (e) .1336
        (c) .1554   (f) .0895

4.7.3. (a) -1.282 (d) 1.282   (g) -0.674
        (b) $\ \ \,0.524$ (e) 0   (h) -0.842
        (c) -0.524 (f) 0   (i) $\ \ \,1.645$

4.7.4. (a) -1.036 (d) -1.405
        (b) 0.674   (e) 2.326
        (c) 0.253   (f) -1.645

4.7.5. (a) $-1.282 < z < 1.282$,   (b) $-2.576 < z < 2.576$.

4.7.6. (a) $-0.674, 0, +0.674$;   (b) $-0.842, -0.253, +0.253, +0.842$.

4.8.1. $\mu = 4.00$, $\sigma = 3.32$.

4.8.2. (a) $\mu = 4$, $\sigma^2 = 5.25$;   (b) $\mu = 1.00$, $\sigma^2 = 8.70$.

4.8.3. $\mu = 2$, $\sigma^2 = 1$.

4.8.4. $\mu = 4.0000$, $\sigma^2 = 2.4002$ (4.0, 2.4).

4.8.5. (a) .6915, (b) .2420, (c) .0548, (d) .7881.

4.8.6.   47, 6.

4.8.7.   a. 0.0062;   b. (a) 1151,   (b) 1587.

4.8.8.   (a) 1.79%,   (b) 84.64%,   (c) 127.3 psi,   (d) 123.4, 136.6 psi.

4.8.9.   a. 82.0%   c. 1.4%, 0.3%
         b. 68.3%, 95.5%, 99.7%   d. 140 or above

4.8.10. 10 to 18 days, 7 to 21 days.

4.9.1.   (a) .1587,   (b) .0179.

4.9.2.   .7119; P(number of deaths $\leq 420 | n = 2000$, $p = .23) = .0179$.

4.9.3.   .3805.

4.9.4.   .0359.

4.9.5.   .2742.

4.9.6.   .2119.

4.9.7. $P(846 \leq Y \leq 954 | Y$ binomial, $n = 5400$, $p = 1/6) = .9546$.

5.3.2.   (a) 343, 294, 418, 446, 471, 691.
         (b) 3275, 0188, 4118, 2591, 2889.
         (c) 044, 099, 053, 025, 513, 339, 484.
         (d) 5298, 3275, 4574, 2297, 1953, 6439, 2214, 2064.

5.3.4.   b. median $= 9600$;   c. mean $= 10,220$.

5.6.1.   a. (i) $14.76 < \mu < 20.24$; (ii) $13.89 < \mu < 21.11$; (iii) 5.48, 7.22.
         b. 2.46, 3.24;   c. 0.78, 1.02;   d. 14.

5.6.2.   a. (i) $220.2 < \mu < 239.8$;   b. 68.

5.6.3. $38.02 < \mu < 39.98$.

5.6.4.   a. $2.91 < \mu < 5.09$;   b. 173.

5.6.5.   151.

5.6.6.   666.

5.6.7.   78.

5.6.8.   Yes: $98.32 < \mu < 129.68$.

5.7.1. $68.53 < \mu < 81.31$.

5.7.2. $43.19 < \mu < 55.81$.

5.7.3.   a. $109.26 < \mu < 138.74$.

5.7.4. $\$52.85 < \mu < \$67.15$.

5.7.5. $7.08 < \mu < 7.32$.

5.7.6. $62.78 < \mu < 65.22$.

5.9.1. a. $-1.13 < \mu_1 - \mu_2 < 2.29$.

5.9.2. a. Coded: $0.17 < \mu_B - \mu_A < 5.41$; in original units: $.04017 < \mu_B - \mu_A < .04541$.

5.9.3. $-6.83 < \mu_A - \mu_B < 2.83$.

5.9.4. a. $-1.24 < \mu_A - \mu_B < 0.52$.

5.9.5. a. $5.76 < \mu < 14.68$; b. $5.15 < \mu < 12.19$; c. 12.97; d. $-3.42 < \mu_1 - \mu_2 < 6.52$.

5.9.6. a. A 95% confidence interval for difference in mean percents of error is $0.26 < \mu_B - \mu_A < 1.48$.

5.9.7. $-0.29 < \mu_U - \mu_T < 11.69$ (U = untreated, T = treated).

5.12.1. a. 10%; b. $2.4\% < p < 17.6\%$.

5.12.2. $49 <$ number of germinating seeds < 111 ($12.2\% < p < 27.8\%$).

5.12.3. $0.392 < p < 0.888$.

5.12.4. $4.1\% < p < 15.9\%$.

5.12.5. $31.1\% < p < 48.9\%$.

5.12.6. $0.00040 < p_A - p_B < 0.05210$.

5.12.7. $0.075 < p_W - p_M < 0.249$.

5.12.8. a. 3%; b. $0.6\% < p < 5.4\%$.

5.12.9. 2655; 956.

5.12.10. b. 84; c. $2.0\% < p < 12.0\%$; d. $24 <$ number defective < 144.

5.12.11. a. 2%; b. $0.9\% < p < 3.1\%$; c. 160 ± 88.

6.6.1. a. $(\bar{Y} - 50)/2 = Z$; reject H_0 if $z \le -1.960$ or if $z \ge +1.960$.
b. $(\bar{Y} - 75)/1.131 = Z$; reject H_0 if $z \ge 2.326$.
c. $(\bar{Y} - 300)/5 = Z$; reject H_0 if $z \le -1.645$.

6.6.2. a. $(\bar{Y} - 50)/2 = Z$; reject H_0 if $z \ge 1.645$.
b. Reject H_0 if $\bar{y} \ge 53.290$. P(accept $H_0 | \mu = 55) = 0.1841$.

6.6.3. $(\bar{Y} - 90)/2.667 = Z$; reject H_0 if $z \le -2.326$.
P (deciding for special study $| \mu = 80) = 0.9192$.

6.6.4. Evidence not sufficient: $z = -1.5$, $P = 0.0668$.

6.6.5. Agree with complaint: $z = 7.00$, $P < 0.00001$.

6.8.1. Accept H_0: $t = 1.20$ (as against critical value 2.539), $.10 < P < .15$.

6.8.2. (a) $\leq .702$ or $\geq .898$,   (b) $\leq .693$ or $\geq .907$,
(c) ≤ 9.67 or ≥ 10.33,   (d) ≤ 9.67 or ≥ 10.33,
(e) ≤ 33.56 or ≥ 46.44,   (f) ≤ 25.40 or ≥ 54.60.

6.8.3. a. Evidence not sufficient: $t = 1.12$ (as against critical value 2.093), $.20 < P < .30$.
  b. Decide against change: $t = 1.12$ (as against critical value -1.729), $.85 < P < .90$.

6.8.4. a. Disagree: $t = -0.534$ (against critical value -2.365), $.50 < P < .70$.
  b. Disagree: $t = 0.814$, $.30 < P < .50$.
  c. $.024664 < \mu < .025212$, $.024857 < \mu < .025293$.
  d. $-.000180 < \mu_b - \mu_a < +.000454$ (95 percent confidence).

6.8.5. a. Not significantly less: $t = -0.586$ (against critical value -1.383), $.25 < P < .35$.

6.9.1. No: $t = 1.33$ (against critical value 2.101), $P = .20$.

6.9.2. a. Yes: $t = 2.39$ (against critical value 1.711), $.01 < P < .025$.
  b. 14.92. $4.24 < \mu_1 - \mu_2 < 25.60$.

6.9.3. Yes: $t = -3.02$; $.001 < P < .01$.

6.9.4. b. No: $t = -0.870$ (against critical value -1.734), $.30 < P < .50$.

6.9.5. Brands as observed differ significantly: $t = 2.45$ (against critical value 2.048), $P \approx .02$; 95 confidence interval is $0.41 < \mu_A - \mu_B < 4.59$.

6.9.6. a. Sample means do not differ significantly: $t = -1.90$ (against critical value -2.179), $.05 < P < .10$.
  b. $-0.41 < \mu_B - \mu_A < 5.99$.

6.9.7. Agree: $t = 3.92$ (against critical value 2.43), $P < .0005$.

6.9.8. a. $4.96 < \mu < 8.38$. b. Reject H_0: $t = 3.41$ (against critical value 3.055), $.001 < P < .01$. c. $0.27 < \mu_a - \mu_b < 5.07$.

6.9.9. a. Observed increase in means is not significant: $t = 0.438$ (against critical value 1.697), $.25 < P < .35$.

6.11.1. Observed rate is significantly lower: $z = -2.50$ (against critical value -2.326), $P = .0062$.

6.11.2. At 1% level, no; at 5% level, yes: $z = 2.10$ (against respective critical values 2.576, 1.960), $P = .0358$.

6.11.3. No: $z = 1.22$ (against critical value 1.960), $P = .23$.

6.11.4. Observed proportion defective not significantly greater than standard: $z = 1.53$ (against critical value 1.645), $P \approx .06$.

6.11.5. No: $z = -2.17$ (against critical value -1.960), $P \approx .03$.

6.11.6. Observed death rate not significantly less than standard: $z = -2.26$ (against critical value -2.326), $P \approx .0107$. [But at any level $\geq .0107$, conclusion would be the opposite.]

6.11.7. Yes: $z = 1.73$ (against critical value 1.645), $P \approx .04$. [But 95% confidence interval is $49.4\% < p < 60.6\%$].

6.11.8. No: 90% confidence interval for p is $45.8\% < p < 62.2\%$, 50% confidence interval is $50.6\% < p < 57.4\%$.

6.11.9. a. No: $z = 0.601$ (against critical value 1.282), $P = .27$.
b. 214, 53.5%.
c. Observed preference not significantly different from 50%: $z = 1.40$ (against critical value 1.960), $P = .16$.

7.8.1. Yes: $\chi_1^2 = 6.628$ versus $\chi_{1*}^2 = 3.841$ ($.01 < P < .025$).

7.8.2. Observed difference not significant: $\chi_1^2 = 0.172$ versus $\chi_{1*}^2 = 3.841$ ($.60 < P < .70$).

7.8.3. b. Accept H_0. $\chi_1^2 = 1.333$ versus $\chi_{1*}^2 = 3.841$ ($.20 < P < .30$).

7.8.4. Evidence not sufficient to reject the hypothesis: (a) $z = -1.130$ versus $z_* = -1.960$; (b) $\chi_1^2 = 1.280$ versus $\chi_{1*}^2 = 3.841$.

7.8.5. Yes: $\chi_1^2 = 109.74$ versus $\chi_{1*}^2 = 3.841$ ($P < .0005$).

7.8.6. No: $\chi_1^2 = 1.905$ versus $\chi_{1*}^2 = 3.841$.

7.8.7. No: $\chi_5^2 = 9.705$ versus $\chi_{5*}^2 = 11.070$.

7.8.8. Reject H_0. $\chi_1^2 = 9.60$ versus $\chi_{1*}^2 = 3.841$.

7.8.9. Observed association is significant: $\chi_4^2 = 11.663$ versus $\chi_{4*}^2 = 9.488$.

7.8.10. Observed association is significant: $\chi_3^2 = 267.81$ versus $\chi_{3*}^2 = 7.815$. In reduced 2×2 table, observed association is significant: $\chi_1^2 = 159.67$ versus $\chi_{1*}^2 = 3.841$.

7.8.11. a. Significantly different: $\chi_2^2 = 58.439$ versus $\chi_{2*}^2 = 5.991$.
b. Significantly different: $\chi_2^2 = 40.646$ versus $\chi_{2*}^2 = 5.991$.

8.3.2. b. $\bar{x} = 9.0$, $\bar{y} = 4.0$.

8.3.3. b. $\bar{x} = 9.0$, $\bar{y} = 3.0$.

8.3.4. a. 1.2; b. $\hat{y} = 4 + 1.2(x - 9)$; c. -6.8.

8.3.5. 0.60, -2.4.

8.3.6. a. 1.6, 2.8, 4.0, 5.2, 6.4; b. -0.6, 0.2, 1.0, -0.2, -0.4.

8.3.7. a. 1.8, 2.4, 3.0, 3.6, 4.2; b. 1.2, -0.4, -2.0, 0.4, 0.8.

8.3.8. a.

	d.f.	S.S.	M.S.
Total (crude)	7	784	
Sample mean	1	700	
Total (corrected for mean)	6	84	
Slope [Regression]	1	75.5712	75.5712
Residual	5	8.4288	1.6858

b. 1.6858; c. 1.30; d. 89.97%.

8.5.1. a. 2.13; b. 3; c. 0.462;
d. accept H_0 ($t = 1.299$ versus $t_* = 3.182$).

8.5.2. $-1.14 < \beta < -.50$. a. 5; b. 2.571; c. 112.

8.8.1. a. 0.491; b. 0.886.

8.8.2. a. 0.886; b. 12.64 and 17.20, 8.74 and 11.26, 2.80 and 7.36.
c. 2.571.

8.8.3. 10.88, 18.96; 6.43, 13.57; 1.04, 9.12.

8.8.4. 0, 4.5; 0.2, 5.4; 1.5, 6.5; 2.6, 7.8; 3.5, 9.3.

8.8.5. $-0.82, 0.27, 1.37, -0.27, -0.55$.

8.8.6. $0.82, -0.27, -1.37, 0.27, 0.55$.

8.9.1. a. $\hat{y} = 7.60 + 0.0629x$.
b. Reject H_0 ($t = 3.51$ versus $t_* = 2.160$)
d. $-.16, -.32, -.29, -.15, .09, .33, .46, .30, .24, .37, .11, -.25,$
$-.41, -.28, -.04$.

8.9.2. Agree: hypothesis test rejects $\beta = 0$ ($t = 2.609$ versus $t_* = 2.447$).

8.9.3. b. $\hat{y} = 592.56 - 6.73x$.
d. yes: $t = -8.00$ versus $t_* = -2.228$.
e. 137.65, 149.01.

8.9.4. a. $\hat{y} = 106.07 - 0.0695x$.
b. -0.0695; c. 0.3651; d. accept H_0 ($t = -0.190$ versus $t_* = -1.714$).

8.9.5. a. $\hat{y} = 0.0334 + 0.7661x$.
b. 0.2147; c. $0.3086 < \beta < 1.2236$; d. 2.9080.

8.9.6. a. 1.26; b. 2.06.

8.9.7. a. $\hat{y} = 50.04 + 0.927x$; b. 927 cases, not significant ($t = 1.38$ versus $t_* = 1.734$).

Index